

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開2003-150418

(P2003-150418A)

(43) 公開日 平成15年5月23日 (2003.5.23)

(51) Int.Cl. <sup>7</sup>	識別記号	F I	テーマコード* (参考)
G 0 6 F 12/00	5 1 4	G 0 6 F 12/00	5 1 4 M 5 B 0 0 5
	5 1 3		5 1 3 D 5 B 0 6 5
3/06	3 0 2	3/06	3 0 2 J 5 B 0 8 2
12/08	5 0 5	12/08	5 0 5 Z

審査請求 未請求 請求項の数18 O L (全 37 頁)

(21) 出願番号 特願2001-345525(P2001-345525)

(22) 出願日 平成13年11月12日 (2001.11.12)

(71) 出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72) 発明者 茂木 和彦

神奈川県川崎市麻生区王禅寺1099番地 株式会社日立製作所システム開発研究所内

(72) 発明者 大枝 高

神奈川県川崎市麻生区王禅寺1099番地 株式会社日立製作所システム開発研究所内

(74) 代理人 100075096

弁理士 作田 康夫

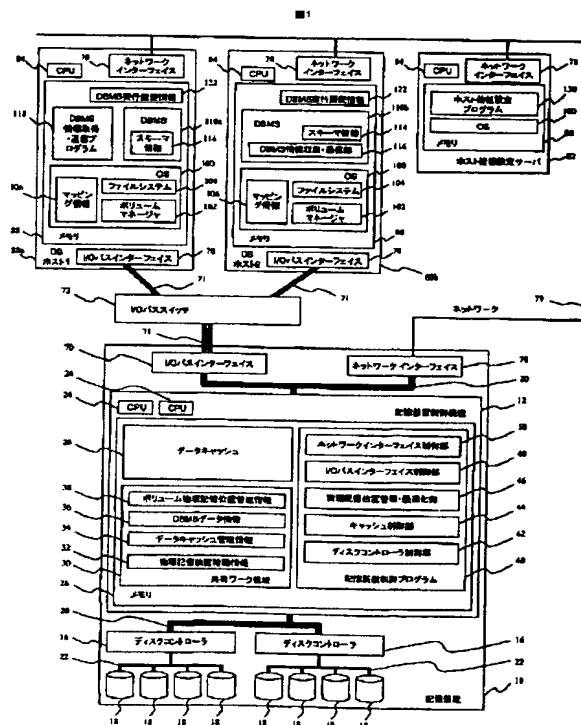
最終頁に続く

(54) 【発明の名称】 データベース管理システムの静的な情報を取得する手段を有する記憶装置

(57) 【要約】

【課題】 データベース管理システム (DBMS) の特性を考慮したデータ配置やキャッシュの制御を行うことにより、記憶装置のデータアクセス性能を向上させる。

【解決手段】 記憶装置は、ネットワーク79を通してDBMSの静的な構成情報をDBMS情報取得・通信プログラムを通して取得し、DBMSデータ情報としてメモリ内に記憶する。記憶装置制御プログラム内の物理記憶位置管理・最適化実行部はDBMSデータ情報を利用してデータ再配置を実行し、キャッシュ制御部はDBMSデータ情報を加味したデータキャッシュ制御を行う。



## 【特許請求の範囲】

【請求項1】データベース管理システムが稼動している計算機との接続手段を有し、

前記データベース管理システムにおけるスキーマにより定義される表・索引・ログを含むデータ構造に関する情報と、前記データベース管理システムが管理するデータベースデータを前記スキーマにより定義されるデータ構造毎に分類した前記記憶装置における記録位置に関する情報を取得する情報取得手段を有することを特徴とする記憶装置。

【請求項2】前記接続手段を用いて複数の前記データベース管理システムが稼動している計算機と接続することを特徴とする請求項1に記載の記憶装置。

【請求項3】前記情報取得手段が複数の前記データベース管理システムが管理するデータベースに関する情報を取得することを特徴とする請求項1に記載の記憶装置。

【請求項4】前記情報取得手段は前記接続手段を用いて情報を取得することを特徴とする請求項1に記載の記憶装置。

【請求項5】前記情報取得手段が、前記データベース管理システムが管理するデータベースに関する情報を前記データベース管理システムから取得することを特徴とする請求項1に記載の記憶装置。

【請求項6】前記情報取得手段が、前記データベース管理システムが管理するデータベースに関する情報を前記データベース管理システムとは異なる少なくとも1つのプログラムを通して取得することを特徴とする請求項1に記載の記憶装置。

【請求項7】前記記憶装置は少なくとも1つ以上のデータを記憶する物理記憶手段を有し、前記計算機が前記記憶装置をアクセスする際に利用する論理的な格納位置を前記記憶装置内で実際にデータを記憶する物理記憶手段の記憶位置へ変換する論理-物理位置変換手段を有し、前記論理位置に対応するデータの物理記憶手段における記憶位置を変更するデータの配置変更手段を有し、前記情報取得手段により取得した情報を利用する前記論理位置に対応するデータの物理記憶位置の変更案を作成する配置変更案作成手段を有することを特徴とする請求項1に記載の記憶装置。

【請求項8】前記配置変更手段を用いて前記配置変更案作成手段により作成されたデータの配置の変更を行なう自動データ再配置手段を有することを特徴とする請求項7に記載の記憶装置。

【請求項9】前記配置変更案作成手段が、前記情報取得手段により取得した情報に基づいて、前記データベース管理システムが前記データベースデータをシーケンシャルにアクセスする際のアクセス場所とアクセス順を判断し、前記シーケンシャルにアクセスされるデータベースデータを前記物理記憶手段上で前記連続アクセス順の前

後関係を崩さずに連続した領域に配置することを特徴とする請求項7に記載の記憶装置。

【請求項10】前記情報取得手段が取得する情報は、前記データベース管理システムが前記スキーマにより定義される同一のデータ構造に属する前記データベースデータをアクセスする際の並列度に関する情報を含み、前記配置変更案作成手段が、前記取得情報を基に、前記スキーマにより定義される同一のデータ構造に属する前記データベースデータを複数の前記物理記憶手段に配置することを特徴とする請求項7に記載の記憶装置。

【請求項11】前記配置変更案作成手段が、前記取得情報を基づいて、同時にアクセスされる可能性が高い前記データベースデータの組を検出し、それらを異なる前記物理記憶手段に配置することを特徴とする請求項7に記載の記憶装置。

【請求項12】前記配置変更案作成手段が、表データと前記表データに対する索引データを抽出し、前記索引が木構造のデータ構造の場合にそれらを前記同時にアクセスされる可能性が高いデータベースデータの組として登録することを特徴とする請求項11に記載の記憶装置。

【請求項13】前記データベースに関する情報に、前記データベース管理システムにおける処理の実行履歴に関する情報を含むことを特徴とする請求項11に記載の記憶装置。

【請求項14】前記物理記憶手段の稼動情報を取得する物理記憶手段稼動情報取得手段を有し、前記配置変更案作成手段が前記物理記憶手段稼動情報取得手段により取得した情報を利用することを特徴とする請求項11に記載の記憶装置。

【請求項15】前記配置変更案作成手段が、前記データベース管理システムにおけるデータの更新処理時に記録するログデータとその他の前記データベースデータを前記同時にアクセスされる可能性が高いデータベースデータの組として登録することを特徴とする請求項11に記載の記憶装置。

【請求項16】キャッシュメモリとデータを記憶する物理記憶手段を有し、前記情報取得手段により取得した情報を利用して前記キャッシュメモリを制御するキャッシュメモリ制御手段を有し、

前記情報取得手段が取得する情報として、前記データベース管理システムが前記計算機上のメモリ上に有するキャッシュの前記データベースデータに対する制御方法とそのキャッシュデータ量に関する情報である計算機キャッシュデータ情報を含むことを特徴とする請求項1に記載の記憶装置。

【請求項17】前記キャッシュメモリ制御手段が、前記スキーマにより定義されるデータ構造に関して前記計算機上メモリにキャッシュされているデータ量である計算機データ構造キャッシュデータ量と前記データ構造の実

データのデータサイズを比較し、前記比較結果を用いて前記データ構造の実データ記憶位置に記憶されているデータのキャッシュ優先度を決定することを特徴とする請求項16に記載の記憶装置。

【請求項18】前記キャッシュメモリ制御手段が、前記計算機データ構造キャッシュデータ量と前記記憶装置における前記データ構造に属するデータに対して前記キャッシュメモリの利用可能量を比較し、前記比較結果を用いて前記データ構造の実データ記憶位置に記憶されているデータのキャッシュ優先度を決定することを特徴とする請求項16に記載の記憶装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、データベース管理システムに関する。

【0002】

【従来の技術】現在、データベース（DB）を基盤とする多くのアプリケーションが存在し、DBに関する一連の処理・管理を行うソフトウェアであるデータベース管理システム（DBMS）は極めて重要なものとなっている。特に、DBMSの処理性能はDBを利用するアプリケーションの性能も決定するため、DBMSの処理性能の向上は極めて重要である。

【0003】DBの特徴の1つは、多大な量のデータを扱うことである。そのため、DBMSの実行環境の多くにおいては、DBMSが実行される計算機に対して大容量の記憶装置を接続し、記憶装置上にDBのデータを記憶する。そのため、DBに関する処理を行う場合に、記憶装置に対してアクセスが発生し、記憶装置におけるデータアクセス性能がDBMSの性能を大きく左右する。そのため、DBMSが稼動するシステムにおいて、記憶装置の性能最適化が極めて重要である。

【0004】文献“Oracle8iパフォーマンスのための設計およびチューニング、リリース8.1”（日本オラクル社、部品番号J00921-01）の第20章（文献1）においては、RDBMSであるOracle8iにおけるI/Oのチューニングについて述べられている。その中で、RDBMSの内部動作のチューニングと共に、データの配置のチューニングに関連するものとして、ログファイルは他のデータファイルから分離した物理記憶装置に記憶すること、ストライプ化されたディスクにデータを記憶することによる負荷分散が効果があること、表のデータとそれに対応する索引データは異なる物理記憶装置に記憶すると効果があること、RDBMSとは関係ないデータを異なる物理記憶装置に記憶することが述べられている。

【0005】米国特許6035306（文献2）においては、DBMS—ファイルシステム—ボリュームマネージャ—記憶装置間のマッピングを考慮した性能解析ツールに関する技術を開示している。この性能解析ツール

は、各レイヤにおけるオブジェクトの稼動状況を画面に表示する。このときに上記のマッピングを考慮し、その各オブジェクトに対応する他レイヤのオブジェクトの稼動状況を示す画面を容易に表示する機能を提供する。また、ボリュームマネージャレイヤのオブジェクトに関して、負荷が高い記憶装置群に記憶されているオブジェクトのうち、2番目に負荷が高いオブジェクトを、もっとも負荷が低い記憶装置群に移動するオブジェクト再配置案を作成する機能を有している。

【0006】特開平9-274544号公報（文献3）においては、計算機がアクセスするために利用する論理的記憶装置を実際にデータを記憶する物理記憶装置上に配置する記憶装置において、前記論理的記憶装置の物理記憶装置への配置を動的に変更することにより記憶装置のアクセス性能を向上する技術について開示している。アクセス頻度が高い物理記憶装置に記憶されているデータの一部を前記の配置動的変更機能を用いて他の物理記憶装置に移動することにより、特定の物理記憶装置のアクセス頻度が高くなるようにし、これにより記憶装置を全体としてみたときの性能を向上させる。また、配置動的変更機能による高性能化処理の自動実行方法についても開示している。

【0007】特開2001-67187号公報（文献4）においては、計算機がアクセスするために利用する論理的記憶装置を実際にデータを記憶する物理記憶装置上に配置し、前記論理的記憶装置の物理記憶装置への配置を動的に変更する機能を有する記憶装置において、論理的記憶装置の物理記憶装置への配置の変更案を作成する際に物理記憶装置を属性の異なるグループに分割し、それを考慮した配置変更案を作成し、その配置変更案に従って自動的に論理的記憶装置の配置を変更する技術について開示している。配置変更案作成時に、物理記憶装置を属性毎にグループ化し、論理的記憶装置の配置先として、それが有している特徴にあった属性を保持している物理記憶装置のグループに属する物理記憶装置を選択する配置変更案を作成することによりより良好なものを作成する。

【0008】米国特許5317727（文献5）においては、DBMSの処理の一部あるいは全部を記憶装置側で実施することによりDBMSの高速化する技術について開示している。記憶装置側でDBMS処理を行うため、記憶装置においてデータアクセス特性を把握することが可能であり、データアクセス特性と記憶装置の構成を考慮することにより無駄な物理記憶媒体に対するアクセスを削減や必要なデータの先読みを実施することができ、結果としてDBMSの性能を向上させることができる。

【0009】論文“高機能ディスクにおけるアクセスプランを用いたプリフェッチ機構に関する評価”（向井他著、第11回データ工学ワークショップ（DEWS20

00) 論文集講演番号3B-3, 2000年7月発行CD-ROM, 主催: 電子情報通信学会データ工学研究専門委員会(文献6)では、記憶装置の高機能化によるDBMSの性能向上について論じている。具体的には、記憶装置に対してアプリケーションレベルの知識としてリレーショナルデータベース管理システム(RDBMS)における問い合わせ処理の実行時のデータのアクセスプランを与えた場合の効果について述べている。更に、その確認のために、トレースデータを用いたホスト側から発行する先読み指示により前記技術を模した擬似実験を実施している。

【0010】

【発明が解決しようとする課題】従来の技術には以下のような問題が存在する。

【0011】文献1で述べられているものは、管理者がデータの配置を決定する際に考慮すべき項目である。現在、1つの記憶装置内に多数の物理記憶装置を有し、多数の計算機により共有されるものが存在する。この記憶装置においては、多くの場合、ホストが認識する論理的記憶装置を実際にデータを記憶する物理記憶装置の適当な領域に割り当てることが行われる。このような記憶装置を利用する場合、人間がすべてを把握することは困難であり、このような記憶装置を含む計算機システム側に何かしらのサポート機能が存在しなければ文献1に述べられている問題点を把握することすら困難となる。また、問題点を把握することができたとしても、計算機システム側にデータの移動機能が存在しない場合には、記憶装置上のデータの再配置のためにデータのバックアップとリストアが必要となり、その処理に多大な労力を必要とする。

【0012】文献2で述べられている技術では、ボリュームマネージャレイヤにおけるオブジェクトの稼動状況による配置最適化案を作成する機能を実現しているが、記憶装置から更に高いアクセス性能を得ようとする場合にはDBMSレイヤにおけるデータの特徴を考慮して配置を決定する必要があるがその点の解決方法に関しては何も述べていない。

【0013】文献3、文献4で述べられている技術に関しては、記憶装置におけるデータ記憶位置の最適化に関する技術である。これらの技術においては記憶装置を利用するアプリケーションが利用するデータに関する特徴としては、アクセス頻度とシーケンシャルアクセス性程度しか考慮していないため、アプリケーションから見た場合に必ずしも最適な配置が実現できるわけではない。例えば、DBMSにおける表データとそれに対応する索引データのような同時にアクセスされるデータを同一の物理記憶装置に配置する可能性がある。このとき、その物理記憶装置においてアクセス競合が発生し、記憶装置のアクセス性能が低下する可能性がある。また、文献1から文献4で述べられている技術に関しては、記憶装置

におけるキャッシュメモリの利用については特に考慮されていない。

【0014】文献5、文献6で述べられている技術に関しては、データの記録位置の最適化に関して特に考慮をしていない。そのため、特定の物理記憶装置の負荷が高くなっているために記憶装置の性能が低下し、DBMSの性能が悪化している状況の解決手段として有効なものではない。

【0015】本発明の第一の目的は、DBMSが管理するデータを保持する記憶装置において、DBMSの処理の特徴を考慮することによりDBMSに対してより好ましい性能特性を持つ記憶装置を実現することである。この記憶装置を利用することにより、既存のDBMSを利用したDBシステムに対しても、DBMSの性能を向上させることができる。

【0016】本発明の第二の目的は、記憶装置の性能最適化機能を提供することにより記憶装置の性能に関する管理コストを削減することである。これにより、この記憶装置を用いたDBシステムのシステム管理コストを削減することができる。

【0017】

【課題を解決するための手段】DBMSに関する情報を記憶装置が取得し、その情報と記憶装置内の物理記憶装置の特性と更に利用可能であれば記憶装置を利用する他のアプリケーションのアクセス頻度に関する情報を考慮する記憶装置の性能最適化処理を記憶装置上で実施する。

【0018】記憶装置がDBMSの特性を考慮してDBMSに良好な性能を得るための手段の第一として、ホストが認識する論理的記憶装置を物理記憶装置の適当な領域に割り当ててデータを記憶する記憶装置における、論理的記憶装置のデータ記憶位置の最適化が存在する。例えば、データ更新時に必ず書き込みが実行される更新ログを、他のデータと異なる物理記憶装置に配置して相互干渉しないようにすることによりDBMSに対して良好な性能特性を得ることができる。

【0019】また、同時にアクセスされる可能性が極めて高い表データとそれに対応する索引データを異なる物理記憶装置に配置することによりDBMSに対して良好な性能特性を得ることができる。更に、DBMSに関する情報を利用して、データがシーケンシャルにアクセスされる場合のアクセス順序を予測し、その構造を保持するように物理記憶装置に記憶することによりシーケンシャルアクセス性能を向上可能である。現在、論理的記憶装置の記憶位置を動的に変更する技術は存在し、これを利用することによりデータの最適配置を実現できる。

【0020】記憶装置がDBMSの特性を考慮してDBMSに良好な性能を得るための手段の第二として、DBMSにおけるホスト上のキャッシュ動作を考慮したキャッシュメモリ制御が存在する。DBMSにおいては利用

頻度の高いデータをホストのメモリ上にキャッシュするが、全てのデータがホストメモリ上に乗ってしまうようなデータに対しては、記憶装置上のキャッシュに保持してもあまり効果はない。

【0021】また、多くのDBMSにおいては、ホスト上のキャッシュの破棄データの選択にLRUアルゴリズムを用いている。ホスト上にキャッシュ可能なデータ量と比べてある一定量以下のデータしか記憶装置上のキャッシュに保持できない場合は、読み出しアクセスにより記憶装置上のキャッシュ上に保持された後にキャッシュに乗っている間に再利用される可能性は低く、そのようなデータを記憶装置上のキャッシュに保持することの効果は小さい。このようなデータをキャッシュから優先的に破棄するような制御を記憶装置上で行うことにより、キャッシュ効果の高いものをより多量に記憶装置のキャッシュメモリ上に保持できるようになり、記憶装置のアクセス性能が向上する。

【0022】

【発明の実施の形態】以下、本発明の実施の形態を説明する。なお、これにより本発明が限定されるものではない。

＜第一の実施の形態＞本実施の形態では、DBMSが実行される計算機と記憶装置が接続された計算機システムにおいて、記憶装置がDBMSに関する情報、記憶装置外におけるデータの記憶位置のマッピングに関する情報を取得し、それらを用いて記憶装置の動作を改善する。記憶装置において、記憶装置内部でデータの記憶位置を動的に変更する機能を有し、取得した情報をもとに好適なデータ再配置案を作成し、データの記憶位置の動的変更機能を用いて、作成したデータ再配置案に従ったデータ配置を実現し、アクセス性能を改善する。また、取得情報をもとにしたデータキャッシュの制御を行いより良いアクセス性能特性が得られるようにする。

【0023】図1は、本発明の第一の実施の形態における計算機システムの構成図である。本実施の形態における計算機システムは、DBホスト80a、80b、ホスト情報設定サーバ82、記憶装置10から構成される。DBホスト80a、80b、ホスト情報設定サーバ82、記憶装置10はそれぞれが保有するネットワークインターフェイス78を通してネットワーク79に接続されている。また、DBホスト80a、80b、記憶装置10はそれぞれが保有するI/Oバスインターフェイス70からI/Oバス71を介してI/Oバススイッチ72に接続され、これらを通して記憶装置10とDBホスト80a、80b間のデータ転送を行う。

【0024】本実施の形態においては、記憶装置10とDBホスト80a、80b間のデータ転送を行うI/Oバス71とネットワーク79を異なるものとしているが、例えばiSCSIのような計算機と記憶装置間のデータ転送をネットワーク上で実施する技術も開発されて

おり、本実施の形態においてもこの技術を利用してもよい。このとき、記憶装置10とDBホスト80a、80bにおいてI/Oバスインターフェイス70が省かれ、計算機システム内からI/Oバス71とI/Oバススイッチ72が省かれる構成となる。

【0025】記憶装置10は、記憶領域を提供するもので、その記憶領域は記憶領域管理単位であるボリュームを用いて外部に提供し、ボリューム内の部分領域に対するアクセスや管理はブロックを単位として実行する。記憶装置10は、ネットワークインターフェイス78、I/Oバスインターフェイス70、記憶装置制御装置12、ディスクコントローラ16、物理記憶装置18から構成され、ネットワークインターフェイス78、I/Oバスインターフェイス70、記憶装置制御装置12、ディスクコントローラ16はそれぞれ内部バス20により接続され、ディスクコントローラ16と物理記憶装置18は物理記憶装置バス22により接続される。記憶装置制御装置12は、CPU24とメモリ26を有する。

【0026】メモリ26上には、記憶装置におけるキャッシュメモリとして利用するデータキャッシュ28が割り当てられ、記憶装置を制御するためのプログラムである記憶装置制御プログラム40が記憶される。また、メモリ26上には、物理記憶装置18の稼動情報である物理記憶装置稼動情報32、データキャッシュ28の管理情報であるデータキャッシュ管理情報34、DBホスト80a、80bで実行されているDBMS110a、110bに関する情報であるDBMSデータ情報36、記憶装置10が提供するボリュームを物理的に記憶する物理記憶装置18上の記憶位置の管理情報であるボリューム物理記憶位置管理情報38を保持する。

【0027】図中の記憶装置10は、複数の物理記憶装置18を有し、1つのボリュームに属するデータを複数の物理記憶装置18に分散配置することが可能である。また、データが記憶される物理記憶装置18上の位置を動的に変更する機能を有する。記憶装置制御プログラム40は、ディスクコントローラ16の制御を行うディスクコントローラ制御部42、データキャッシュ28の管理を行うキャッシュ制御部44、記憶装置10が提供するボリュームを物理的に記憶する物理記憶装置18上の記憶位置の管理やデータが記憶される物理記憶装置18上の位置を動的に変更する機能に関する処理を行う物理記憶位置管理・最適化部46、I/Oバスインターフェイス70の制御を行うI/Oバスインターフェイス制御部48、ネットワークインターフェイス78の制御を行うネットワークインターフェイス制御部50を含む。

【0028】DBホスト80a、80b、ホスト情報設定サーバ82においては、それぞれCPU84、ネットワークインターフェイス78、メモリ88を有し、メモリ88上にオペレーティングシステム(OS)100が記憶・実行されている。

【0029】DBホスト80a, 80bはI/Oバスインターフェイス70を有し、記憶装置10が提供するボリュームに対してアクセスを実行する。OS100内にファイルシステム104と1つ以上のボリュームからホストが利用する論理的なボリュームである論理ボリュームを作成するボリュームマネージャ102と、ファイルシステム104やボリュームマネージャ102により、OS100によりアプリケーションに対して提供されるファイルや論理ローボリュームに記憶されたデータの記録位置等を管理するマッピング情報106を有する。OS100が認識するボリュームやボリュームマネージャ102により提供される論理ボリュームに対して、アプリケーションがそれらのボリュームをファイルと等価なインターフェイスでアクセスするための機構であるローデバイス機構をOS100が有していても良い。

【0030】図中の構成ではボリュームマネージャ102が存在しているが、本実施の形態においてはボリュームマネージャ102における論理ボリュームの構成を変更することはないので、ボリュームマネージャ102が存在せずにファイルシステムが記憶装置10により提供されるボリュームを利用する構成に対しても本実施の形態を当てはめることができる。

【0031】DBホスト80a, 80bのそれぞれのメモリ88上ではDBMS110a, 110bが記憶・実行され、実行履歴情報122が記憶されている。DBMS110a, 110bは内部にスキーマ情報114を有している。図中では、DBMS110a, 110bが1台のホストに1つのみ動作しているが、後述するように、DBMS110a, 110b毎の識別子を用いて管理を行うため、1台のホストにDBMSが複数動作していても本実施の形態にあてはめることができる。

【0032】DBホスト80a上ではDBMS情報取得・通信プログラム118が動作している。一方、DBホスト80b上ではDBMS情報取得・通信プログラム118が提供する機能をDBMS110b中のDBMS情報収集・通信部116が提供する。

【0033】ホスト情報設定サーバ82のメモリ88上ではホスト情報設定プログラム130が記憶・実行される。

【0034】図2はDBホスト80a, 80bのOS100内に記憶されているマッピング情報106を示す。マッピング情報106中には、ボリュームローデバイス情報520、ファイル記憶位置情報530と論理ボリューム構成情報540が含まれる。

【0035】ボリュームローデバイス情報520中にはOS100においてローデバイスを指定するための識別子であるローデバイスパス名521とそのローデバイスによりアクセスされる記憶装置10が提供するボリュームあるいは論理ボリュームの識別子であるローデバイスボリューム名522の組が含まれる。

【0036】ファイル記憶位置情報530中には、OS100においてファイルを指定するための識別子であるファイルパス名531とそのファイル中のデータ位置を指定するブロック番号であるファイルブロック番号532とそれに対応するデータが記憶されている記憶装置10が提供するボリュームもしくは論理ボリュームの識別子であるファイル配置ボリューム名533とそのボリューム上のデータ記憶位置であるファイル配置ボリュームブロック番号534の組が含まれる。

【0037】論理ボリューム構成情報540中にはボリュームマネージャ102により提供される論理ボリュームの識別子である論理ボリューム名541とその論理ボリューム上のデータの位置を示す論理ボリューム論理ブロック番号542とその論理ブロックが記憶されているボリュームの識別子であるボリューム名501とボリューム上の記憶位置であるボリューム論理ブロック番号512の組が含まれる。マッピング情報106を取得するには、OS100が提供している管理コマンドの実行や情報提供機構の利用、場合によっては参照可能な管理データの直接解析等を行う必要がある。

【0038】図3はDBMS110a, 110b内に記憶されているその内部で定義・管理しているデータその他の管理情報であるスキーマ情報114を示す。スキーマ情報114には、表のデータ構造や制約条件等の定義情報を保持する表定義情報551、索引のデータ構造や対象である表等の定義情報を保持する索引定義情報552、利用するログに関する情報であるログ情報553、利用する一時表領域に関する情報である一時表領域情報554、管理しているデータのデータ記憶位置の管理情報であるデータ記憶位置情報555、キャッシュの構成に関する情報であるキャッシュ構成情報556とデータをアクセスする際の並列度に関する情報である最大アクセス並列度情報557を含む。

【0039】データ記憶位置情報555中には、表、索引、ログ、一時表領域等のデータ構造の識別子であるデータ構造名561とそのデータを記憶するファイルまたはローデバイスの識別子であるデータファイルパス名562とその中の記憶位置であるファイルブロック番号563の組が含まれる。キャッシュ構成情報556はDBMS110a, 110bが三種類のキャッシュ管理のグループを定義し、そのグループに対してキャッシュを割り当てている場合を示す。

【0040】キャッシュ構成情報556中には、グループ名565とグループ中のデータ構造のデータをホスト上にキャッシュする際の最大データサイズであるキャッシュサイズ566とそのグループに所属するデータ構造の識別子の所属データ構造名567の組が含まれる。最大アクセス並列度情報557には、データ構造名561とそのデータ構造にアクセスする際の一般的な場合の最大並列度に関する情報である最大アクセス並列度569

の組が含まれる。

【0041】スキーマ情報114を外部から取得するには、管理ビューとして外部に公開されているものをSQL等のデータ検索言語を用いて取得したり、または、専用の機構を用いて取得したりすることができる。

【0042】図4はDBホスト80a、80bのメモリ88上に記憶されている実行履歴情報122を示す。実行履歴情報122中には、DBMS110a、110bで実行されたクエリ570の履歴が記憶されている。この情報は、DBMS110a、110bで作成する。またはDBMSのフロントエンドプログラムがこの情報を作成する。この場合には、DBMSフロントエンドプログラムが存在する計算機に実行履歴情報122が記憶されることになる。

【0043】図5は記憶装置10内に保持されているボリューム物理記憶位置管理情報38を示す。ボリューム物理記憶位置管理情報38中には、データの論理アドレス物理記憶装置18における記憶位置のマッピングを管理するボリューム物理記憶位置メイン情報510と記憶装置10内でのボリュームに属するデータの物理記憶位置の変更処理の管理情報であるボリュームデータ移動管理情報511が含まれる。

【0044】ボリューム物理記憶位置メイン情報510中には、ボリューム名501とそのボリューム上のデータ記憶位置であるボリューム論理ブロック番号512とその論理ブロックが記憶されている物理記憶装置18の識別子である物理記憶装置名502と物理記憶装置18上の記憶位置である物理ブロック番号514の組のデータが含まれる。ここで、ボリューム名501が“Empty”であるエントリ515は特殊なエントリであり、このエントリには記憶装置10内の物理記憶装置18の領域のうち、ボリュームに割り当てられていない領域を示し、この領域に対してデータをコピーすることによりデータの物理記憶位置の動的変更機能を実現する。

【0045】ボリュームデータ移動管理情報511はボリューム名501と、そのボリューム内の記憶位置を変更するデータ範囲を示す移動論理ブロック番号782と、そのデータが新規に記憶される物理記憶装置18の識別子とその記憶領域を示す移動先物理記憶装置名783と移動先物理ブロック番号784、現在のデータコピー元を示すコピーポインタ786とデータの再コピーの必要性を管理する差分管理情報785の組が含まれる。

【0046】差分管理情報785とコピーポインタ786を用いたデータの記憶位置変更処理の概略を以下に示す。差分管理情報785はある一定量の領域毎にデータコピーが必要である「1」または不必要「0」を示すデータを保持する。データの記憶位置変更処理開始時に全ての差分管理情報785のエントリを1にセットし、コピーポインタ786を移動元の先頭にセットする。コピーポインタ786にしたがって差分管理情報785に1

がセットされている領域を順次移動先にデータをコピーし、コピーポインタ786を更新していく。差分管理情報785で管理される領域をコピーする直前に、その対応するエントリを0にセットする。

【0047】データコピー中に移動領域内のデータに対する更新が行われた場合、それに対応する差分管理情報785のエントリを1にセットする。一度全領域のコピーが完了した段階で差分管理情報785内のエントリが全て0になったかを確認し、全て0であればボリューム物理記憶位置メイン情報510を更新してデータの記憶位置変更処理は完了する。1のエントリが残っている場合には、再度それに対応する領域をコピーする処理を前記手順で繰り返す。

【0048】なお、データ記憶位置の動的変更機能の実現方法は他の方式を用いても良い。この場合には、ボリューム物理記憶位置管理情報38中にはボリュームデータ移動管理情報511ではなく他のデータ記憶位置の動的変更機能のための管理情報が含まれることになる。

【0049】図6に記憶装置10内に保持されている物理記憶装置稼働情報32を示す。物理記憶装置稼働情報32中には、記憶装置10が提供するボリュームの識別子であるボリューム名501とそのボリューム名501を持つボリュームのデータを保持する物理記憶装置18の識別子である物理記憶装置名502、そしてボリューム名501を持つボリュームが物理記憶装置名502を持つ物理記憶装置18に記憶しているデータをアクセスするための稼働時間のある時刻からの累積値である累積稼働時間503、稼働率594計算のために前回利用した累積稼働時間503の値である旧累積稼働時間593とある一定時間内の動作時間の割合を示す稼働率594の組と、稼働率594計算のために前回累積稼働時間取得時刻595を含む。

【0050】ディスクコントローラ制御部42はディスクコントローラ16を利用して物理記憶装置18へのデータアクセスする際の開始時刻と終了時刻を取得し、そのアクセスデータがどのボリュームに対するものかを判断して開始時刻と終了時刻の差分を稼働時間として対応するボリューム名501と物理記憶装置名502を持つデータの組の累積稼働時間503に加算する。

【0051】物理記憶位置管理・最適化部46は一定間隔で以下の処理を行う。累積稼働時間503と旧累積稼働時間593、前回累積稼働時間取得時刻595と現データ取得時刻を用いて前回累積稼働時間取得時刻595と現データ取得時刻間の稼働率594を計算・記憶する。その後、取得した累積稼働時間503を旧累積稼働時間593に、現データ取得時刻を前回累積稼働時間取得時刻595に記憶する。

【0052】図7に記憶装置10内に保持されているDBMSデータ情報36を示す。DBMSデータ情報36

中には、DBMSスキーマ情報711、データ構造物理記憶位置情報712、DBMS実行履歴情報714、DBMSデータ構造キャッシュ効果情報715を含む。

【0053】DBMSデータ情報36中に含まれるデータは、DBホスト80a、80b上に存在するデータを利用する必要があるものが含まれる。記憶装置10は記憶装置10の外部に存在する情報をホスト情報設定サーバ82で動作するホスト情報設定プログラム130を利用して取得する。ホスト情報設定プログラム130はネットワーク79を通し、DBホスト80a上で実行され、マッピング情報106等必要となる情報の収集処理を実施するDBMS情報取得・通信プログラム118や、DBホスト80b上で実行されているDBMS110b中のDBMS情報取得・通信プログラム118と等価な機能を実現するDBMS情報収集・通信部116を利用して必要な情報を収集する。

【0054】ホスト情報設定プログラム130は情報取得後、必要ならば記憶装置10に情報を設定するためのデータの加工を行い、ネットワーク79を通して記憶装置10に転送する。記憶装置10においては、ネットワークインターフェイス制御部50が必要な情報が送られてきたことを確認し、物理記憶位置管理・最適化部46に渡し、必要な加工を行った後にその情報をDBMSデータ情報36中の適切な場所に記憶する。

【0055】ホスト情報設定プログラム130は任意のDBホスト80a、80b上で実行されていてもよい。あるいは、物理記憶位置管理・最適化部46がホスト情報設定プログラム130の情報収集機能を有してもよい。これらの場合は、DBホスト80a、80bから情報を転送する際にI/Oパス71を通して行ってもよい。この場合、特定の領域に対する書き込みが特定の意味を持つ特殊なボリュームを記憶装置10はDBホスト80a、80bに提供し、そのボリュームに対する書き込みがあった場合にI/Oパスインターフェイス制御部70は情報の転送があったと判断し、その情報を物理記憶位置管理・最適化部46に渡し、必要な加工を行った後にその情報をDBMSデータ情報36中の適切な場所に記憶する、等の方式を利用する。

【0056】情報の収集処理に関しては、記憶装置10が必要になったときに外部にデータ転送要求を出す方法と、データの変更があるたびに外部から記憶装置10に変更されたデータを送る方法の2種類ともに利用することができる。

【0057】図8にDBMSデータ情報36中に含まれるDBMSスキーマ情報711を示す。DBMSスキーマ情報711は、DBMSデータ構造情報621、DBMSデータ記憶位置情報622、DBMSパーティション化・索引情報623、DBMS索引定義情報624、DBMSキャッシュ構成情報625、DBMSホスト情報626を含む。

【0058】DBMSデータ構造情報621はDBMS110a、110bで定義されているデータ構造に関する情報で、DBMS110a、110bの識別子であるDBMS名631、DBMS110a、110b内の表・索引・ログ・一時表領域等のデータ構造の識別子であるデータ構造名561、データ構造の種別を表すデータ構造種別640、データ記憶位置情報から求めることができるデータ構造が利用する総データ量を示すデータ構造データ量641、そのデータ構造をアクセスする際の最大並列度に関する情報である最大アクセス並列度569の組を保持する。このとき、データ構造によっては最大アクセス並列度569の値を持たない。

【0059】DBMSデータ記憶位置情報622はDBMS名631とそのDBMSにおけるデータ記憶位置管理情報555であるデータ記憶位置管理情報638の組を保持する。

【0060】DBMSパーティション化表・索引情報623は、1つの表や索引をある属性値により幾つかのグループに分割したデータ構造を管理する情報で、パーティション化されたデータ構造が所属するDBMS110a、110bの識別子であるDBMS名631と分割化される前のデータ構造の識別子であるパーティション元データ構造名643と分割後のデータ構造の識別子であるデータ構造名561とその分割条件を保持するパーティション化方法644の組を保持する。今後、パーティション化されたデータ構造に関しては、特に断らない限り単純にデータ構造と呼ぶ場合にはパーティション化後のものを指すものとする。

【0061】DBMS索引定義情報624には、DBMS名631、索引の識別子である索引名635、その索引のデータ形式を示す索引タイプ636、その索引がどの表のどの属性に対するものかを示す対応表情報637の組を保持する。

【0062】DBMSキャッシュ構成情報625は、DBMS110a、110bのキャッシュに関する情報であり、DBMS名631とDBMS110a、110bにおけるキャッシュ構成情報556の組を保持する。

【0063】DBMSホスト情報626は、DBMS名631を持つDBMS110a、110bがどのホスト上で実行されているかを管理するもので、DBMS名631とDBMS実行ホストの識別子であるホスト名651の組を保持する。DBMSスキーマ情報711中のDBMSホスト情報626以外は、DBMS110a、110bが管理しているスキーマ情報114の中から必要な情報を取得して作成する。DBMSホスト情報626はシステム構成情報で管理者が設定するものである。

【0064】図9にDBMSデータ情報36中に含まれるデータ構造物理記憶位置情報712を示す。データ構造物理記憶位置情報712はDBMS110a、110bに含まれるデータ構造が記憶装置10内でどの物理記



憶装置18のどの領域に記憶されるかを管理するもので、データ構造を特定するDBMS名631とデータ構造名561、その外部からのアクセス領域を示すボリューム名501とボリューム論理ブロック番号512、その物理記憶装置18上の記憶位置を示す物理記憶装置名502と物理ブロック番号514の組を保持する。この情報は、DBMSデータ記憶位置情報622とマッピング情報106を記憶装置10の外部から取得し、さらにボリューム物理記憶位置メイン情報510を参照して、対応する部分を組み合わせることにより作成する。

【0065】DBMS110a, 110b毎にシーケンシャルアクセスの方法が定まっている。データ構造物理記憶位置情報712を作成する際に、DBMS名631とデータ構造名561により特定されるデータ構造毎に、シーケンシャルアクセス時のアクセス順を保持するようにソートしたデータを作成する。ここでは、対象とするDBMS110a, 110bの種類を絞り、あらかじめデータ構造物理記憶位置情報712を作成するプログラムがDBMS110a, 110bにおけるシーケンシャルアクセス方法を把握し、シーケンシャルアクセス順でソートされたデータを作成する。

【0066】本実施の形態のDBMS110a, 110bにおけるシーケンシャルアクセス方法は以下の方法に従うものとする。あるデータ構造のデータをシーケンシャルアクセスする場合に、データ構造が記憶されているデータファイル名562とファイルブロック番号563を昇順にソートしその順序でアクセスを実行する。その他にシーケンシャルアクセス方法の決定方法としては、データファイルを管理する内部通番とファイルブロック番号563の組を昇順にソートした順番にアクセスする方法等が存在し、それらを利用したシーケンシャルアクセス方法の判断を実施してもよい。

【0067】図10にDBMSデータ情報36中に含まれるクエリ実行同時アクセスデータ構造カウント情報714を示す。これは、実行履歴情報122をもとに同時にアクセスされるデータ構造の組と実行履歴中に何回その組を同時にアクセスするクエリが実行されたかを示すデータで、DBMS名631、同時にアクセスされる可能性のあるデータ構造のデータ構造名561の組を示すデータ構造名A701とデータ構造名B702、そして、DBMS実行履歴122の解析によりそのデータ構造の組がアクセスされたと判断された回数であるカウント値703の組で表される。この組はカウント値703の値でソートしておく。

【0068】クエリ実行時同時アクセスデータカウント情報714はDBMS実行履歴122から作成する。最初にクエリ実行時同時アクセスデータカウント情報714のエントリを全消去する。DBMS110a, 110bにおいて定型処理が行われる場合には、まず、その型により分類し、その型の処理が何回実行されたかを確認

する。続いてDBMS110a, 110bから型毎のクエリ実行プランを取得する。そのクエリ実行プランにより示される処理手順から同時にアクセスされるデータ構造の組を判別する。

【0069】そして、クエリ実行時同時アクセスデータカウント情報714中のDBMS名631・データ構造名A701・データ構造名B702を参照し、既に対応するデータ構造の組が存在している場合には先に求めたその型の処理回数をカウント値703に加算する。既に対応するデータ構造の組が存在していない場合には、新たにエントリを追加してカウント値703を先に求めたその型の処理回数にセットする。

【0070】DBMS110a, 110bにおいて非定型処理が行われる場合には、1つ1つの実行されたクエリに関してクエリ実行プランを取得し、そのクエリ実行プランにより示される処理手順から同時にアクセスされるデータ構造の組を判別する。そして、クエリ実行時同時アクセスデータカウント情報714中のDBMS名631・データ構造名A701・データ構造名B702を参照し、既に対応するデータ構造の組が存在している場合にはカウント値703に1を加算する。既に対応するデータ構造の組が存在していない場合には、新たにエントリを追加してカウント値703に1をセットする。

【0071】クエリ実行プランから同時にアクセスされる可能性があるデータ構造の判別は以下に行う。まず、木構造の索引に対するアクセスが実施される場合には、その木構造索引データと、その索引が対象とする表データが同時にアクセスされると判断する。また、データの更新処理や挿入処理が行われる場合には、ログとその他のデータが同時にアクセスされると判断する。

【0072】以下はDBMS110a, 110bの特性に依存するが、例えば、クエリ実行プラン作成時にネストループジョイン処理を多段に渡り実行する計画を作成し、それらの多段に渡る処理を同時に実行するRDBMSが存在する。このRDBMSを利用する場合にはその多段に渡るネストループジョイン処理で利用する表データとその表に対する木構造の索引データは同時にアクセスされると判断できる。

【0073】このように、クエリ実行計画による同時アクセスデータの判断に関しては、DBMS110a, 110bの処理特性を把握して判断する必要があるが、ここでは、対象とするDBMS110a, 110bの種類を絞り、クエリ実行時同時アクセスデータカウント情報714を作成するプログラムがDBMS110a, 110b特有の同時アクセスデータ構造の組を把握できる機能を有することを仮定する。

【0074】実行履歴情報122からクエリ実行時同時アクセスデータカウント情報714を作成する処理は、記憶装置10の内部、外部どちらで実行してもよい。記憶装置10でクエリ実行時同時アクセスデータカウント

情報714を作成する場合には、記憶装置10がネットワーク79を通してDBホスト80a、80b、あるいは、実行履歴情報122がDBMSフロントエンドプログラムが実行される計算機上に記憶される場合にはその計算機に対して実行履歴情報122を記憶装置10に転送する要求を出し、その情報をネットワーク79を通して受け取る。

【0075】その後、前述のクエリ実行時同時アクセスデータカウント情報714作成処理を実施する。記憶装置10の外部で作成する場合は、例えば、ホスト情報設定サーバ82がDBホスト80a、80b、あるいはDBMSフロントエンドプログラムが実行される計算機から実行履歴情報122を取得し、クエリ実行時同時アクセスデータカウント情報714作成処理を実施する。その後、ネットワーク79を通して作成されたクエリ実行時同時アクセスデータカウント情報714を記憶装置10に転送し、それをDBMSデータ情報36中に記憶する。

【0076】なお、本実施の形態においては、常に実行履歴情報122が作成される必要はない。クエリ実行時同時アクセスデータカウント情報714作成時に実行履歴情報122が存在しないDBMS110a、110bが利用するデータ構造に関してはそれらを見捨ててデータを作成する。また、クエリ実行時同時アクセスデータカウント情報714は存在しなくてもよい。

【0077】図11にDBMSデータ情報36に含まれるDBMSデータ構造キャッシュ効果情報715を示す。DBMSデータ構造キャッシュ効果情報715は記憶装置10においてデータ構造をデータキャッシュに保持しておくことに効果があるかどうかを判断した結果を保持するもので、データ構造を特定するDBMS名631とデータ構造名561、そのデータ構造がデータキャッシュに保持する効果があるかどうかの判断結果を示すキャッシュ効果情報733を保持する。キャッシュ効果情報733の値は、管理者が指定する、もしくは、以下に述べる手順に従って求めるものである。

【0078】図12に記憶装置10において指定されたデータ構造のデータをデータキャッシュに保持する効果があるかどうかの判断する処理のフローを示す。判断基準は2種類有り、1つは「指定データ構造のデータ量に比べてホストキャッシュ量が十分に存在するために利用頻度が高いデータの読出しアクセスが実行されないか」で、もう1つは「記憶装置10のデータキャッシュ量がホストキャッシュ量に比べて小さく、記憶装置10のデータキャッシュ量で効果がある利用頻度のデータはホストキャッシュに載ってしまい、記憶装置10から読出されるデータを記憶装置10でキャッシュしても効果が低い」ことである。

【0079】ステップ2801で処理を開始する。ステップ2802で指定データ構造と同じキャッシュ管理の

グループに属するデータ構造のデータ量の総和をDBMSキャッシュ構成情報625とDBMSデータ構造情報621を参照して求める。

【0080】ステップ2803で指定データ構造と同じキャッシュ管理のグループにおけるそのグループの単位データ量あたりのホストにおける平均キャッシュ量を前記のグループのデータ総量とDBMSキャッシュ構成情報625中のキャッシュサイズ566から求め、その値をあらかじめ定められたキャッシュ効果判断閾値と比較する。その値が閾値以上の場合にはステップ2807に進み、閾値未満の場合にはステップ2804に進む。この閾値としては概ね0.7程度の値を用いる。

【0081】ステップ2804では記憶装置10における単位容量あたりの平均キャッシュデータ量を求める。この値は、記憶装置のデータキャッシュ28の総容量と外部に提供するボリュームの総容量から求めることができ、これらの値はボリューム物理記憶位置管理情報38やデータキャッシュ管理情報34を参照することにより求めることができる。

【0082】ステップ2805では、前述の指定データ構造が属するキャッシュ管理のグループにおける単位データ量あたりのホストにおける平均キャッシュ量に対する記憶装置10における平均キャッシュ量の比率を求め、その値がキャッシュ効果判断閾値未満の場合はステップ2807に進み、閾値以上の場合にはステップ2806に進む。この閾値としては概ね0.7程度の値を用いる。

【0083】ステップ2806では記憶装置10においてキャッシュする効果があると判定し、ステップ2808に進みキャッシュ効果判定処理を終了する。

【0084】ステップ2807では記憶装置10においてキャッシュする効果がないと判定し、ステップ2808に進みキャッシュ効果判定処理を終了する。

【0085】記憶装置10は、データキャッシュをある一定サイズの領域であるセグメントと呼ぶ管理単位を用いて管理する。図13に記憶装置10内に保持されているデータキャッシュ管理情報34を示す。データキャッシュ管理情報34中には、データキャッシュ28のセグメントの状態を示すキャッシュセグメント情報720とキャッシュセグメントの再利用対象選定に利用するキャッシュセグメント利用管理情報740を含む。

【0086】キャッシュセグメント情報720中には、セグメントの識別子であるセグメントID721と、そのセグメントに記憶されているデータ領域を示すボリューム名511とボリューム論理ブロック番号512、そして、セグメントの状態を示すステータス情報722、後述するセグメントの再利用選定管理に利用するリストの情報であるリスト情報723を含む。

【0087】ステータス情報722が示すセグメントの状態としては、物理記憶装置18上にセグメント内のデ

ータと同じデータが記憶されている“ノーマル”、セグメント内にのみ最新のデータが存在する“ダーティ”、セグメント内に有効なデータが存在しない“インバリッド”が存在する。リスト情報723には、現在そのセグメントが属するリストの識別子と、そのリストのリンク情報が記憶される。図中では、リストは双方向リンクリストであるとしている。

【0088】キャッシュセグメント利用管理情報740中には、キャッシュセグメントの再利用対象選定に利用する3種類の管理リストである第1LRUリスト、第2LRUリスト、再利用LRUリストの管理情報として、第1LRUリスト情報741、第2LRUリスト情報742、再利用LRUリスト情報743が記憶される。

【0089】第1LRUリスト情報741、第2LRUリスト情報742、再利用LRUリスト情報743は、それぞれリストの先頭であるMRUセグメントID、最後尾であるLRUセグメントID、そのリストに属するセグメント数を記憶する。この3種類の管理リストはホストからのアクセス要求の処理にかかわるもので、アクセス要求処理の説明時に同時に行う。

【0090】ホストからのデータアクセス要求があったときの処理を説明する。

【0091】図14に記憶装置10がホストからデータの読出し要求を受け取ったときの処理フローを示す。ステップ2901で、I/Oバスインターフェイス70はホストからのデータ読出し要求を受け、I/Oバスインターフェイス制御部48がその要求を認識する。

【0092】ステップ2902でキャッシュ制御部44は読出し要求があったデータがデータキャッシュ28上に存在するかデータキャッシュ管理情報34を参照して確認する。存在する場合にはステップ2905に進み、存在しない場合にはステップ2903に進む。

【0093】ステップ2903で、キャッシュ制御部44は読出し要求があったデータを保持するキャッシュ領域を確保する。データを保持するキャッシュセグメントとして、ステータス情報がノーマルのもののうち、再利用LRUリストのLRU (Least Recently Used: 最も昔に使われた) 側に存在するものを必要数取得し、再利用LRUリストから削除する。そして、再利用LRUリスト情報743をそれに合わせて更新する。また、キャッシュセグメント情報720中のボリューム名511とボリューム論理ブロック番号を記憶するデータのものと変更し、ステータス情報722をインバリッドに設定する。

【0094】ステップ2904でディスクコントローラ制御部42は読出し要求があったデータを物理記憶装置18から読み出す処理を実施し、その完了を待つ。読出し完了後、キャッシュセグメント情報720中の対応するステータス情報722をノーマルに設定し、ステップ2906に進む。

【0095】ステップ2905でキャッシュ制御部44はデータ読出し要求のあったデータを保持するセグメントをその管理のためにリンクされている管理リストから削除する。

【0096】ステップ2906でI/Oバスインターフェイス管理部48はデータ読出し要求のあったデータをセグメントからI/Oバスインターフェイス70を利用してホストに転送し、ホストとの処理を完了する。

【0097】ステップ2907でキャッシュ制御部44はアクセス先のデータの内容に従い、データ読出し要求のあったデータを保持するセグメントを適当な管理リストに繋ぐ処理を行う。この処理の詳細は後述する。

【0098】ステップ2908でホストからの読出し要求を受けとった時の処理を終了する。

【0099】図15に記憶装置10がホストからデータの書き込み要求を受け取ったときの処理フローを示す。ステップ2931で、I/Oバスインターフェイス70はホストからのデータ書き込み要求を受け、I/Oバスインターフェイス制御部48がその要求を認識する。

【0100】ステップ2932でキャッシュ制御部44は読出し要求があったデータを保持するセグメントがデータキャッシュ28上に存在するかデータキャッシュ管理情報34を参照して確認する。存在する場合にはステップ2934に進み、存在しない場合にはステップ2933に進む。

【0101】ステップ2933で、キャッシュ制御部44は書き込み要求があったデータを保持するキャッシュ領域を確保する。データを保持するキャッシュセグメントとして、ステータス情報がノーマルのもののうち、再利用LRUリストのLRU側に存在するものを必要数取得し、再利用LRUリストから削除する。そして、再利用LRUリスト情報743をそれに合わせて更新する。また、キャッシュセグメント情報720中のボリューム名511とボリューム論理ブロック番号を記憶するデータのものと変更し、ステータス情報722をインバリッドに設定する。

【0102】ステップ2934でキャッシュ制御部44はデータ書き込み要求のあったデータを保持するセグメントをその管理のためにリンクされている管理リストから削除する。

【0103】ステップ2935でI/Oバスインターフェイス管理部48はデータ書き込み要求のあったデータをキャッシュセグメントに書き込み、キャッシュセグメント情報720中の対応するステータス情報722をダーティに設定し、ホストとの処理を完了する。

【0104】ステップ2936でキャッシュ制御部44はアクセス先のデータの内容に従い、データ書き込み要求のあったデータを保持するセグメントを適当な管理リストに繋ぐ処理を行う。この処理の詳細は後述する。

【0105】ステップ2937でホストからの書き込み要

求を受けとった時の処理を終了する。

【0106】図16にキャッシュ制御部44が実行するアクセス先のデータの内容に従い、アクセス要求のあったデータを保持するセグメントを適当な管理リストに繋ぐ処理のフローを示す。この処理において、記憶装置10におけるキャッシュ効果がないと判断されるデータを保持するキャッシュセグメントを管理リスト中の再利用されやすい場所に繋ぐことによりキャッシュ効果がないと判断されるものがデータキャッシュ28上に載っている時間を短くし、他のデータのキャッシュ効果を高めることを行う。

【0107】ステップ2961においてアクセス先のデータの内容に従い、アクセス要求のあったデータを保持するセグメントを適当な管理リストに繋ぐ処理を開始する。

【0108】ステップ2962において、アクセス先データのキャッシュ効果の確認を行う。データ構造物理記憶位置情報712を参照してアクセス先データが属するDBMS110a、110bとそのデータ構造の識別子であるDBMS名631とデータ構造名561を求める。データ構造物理記憶位置情報712に対応部分がない場合にはキャッシュ効果があると判断する。

【0109】続いて、DBMSデータ構造キャッシュ効果情報715を参照し、既に求めたDBMS名631とデータ構造名561に対応するキャッシュ効果情報733を参照し、アクセス先データにキャッシュ効果があるかないかを求める。なお、キャッシュ効果情報733中に対応するエントリがない場合、キャッシュ効果があると判断する。キャッシュ効果があると判断された場合にはステップ2963に進み、ないと判断された場合にはステップ2966に進む。

【0110】ステップ2963でアクセス先データを保持するキャッシュセグメントを第1LRUリストのMRU(Most Recently Used:最も最近使われた)側にリンクし、それに合わせて第1LRUリスト情報741を更新する。

【0111】ステップ2964で第1LRUリストにリンクされているセグメント数を第1LRUリスト情報741を参照して確認し、その値が事前に定めてある閾値を超えているか確認する。そのセグメント数が閾値未満の場合にはステップ2970に進み処理を完了する。閾値以上の場合にはステップ2965に進む。

【0112】ステップ2965で第1LRUリストのセグメント数が閾値未満になるように第1LRUの最もLRU側に存在するセグメントを第2LRUリストのMRU側にリンクし直す処理を行い、それに合わせて第1LRUリスト情報741と第2LRUリスト情報742を更新し、ステップ2967に進む。

【0113】ステップ2966でアクセス先データを保持するキャッシュセグメントを第2LRUリストのMR

U側にリンクし、それに合わせて第2LRUリスト情報742を更新し、ステップ2967に進む。

【0114】ステップ2967で第2LRUリストにリンクされているセグメント数を第2LRUリスト情報742を参照して確認し、その値が事前に定めてある閾値を超えているか確認する。そのセグメント数が閾値未満の場合にはステップ2970に進み処理を完了する。閾値以上の場合にはステップ2968に進む。

【0115】ステップ2968で第2LRUリストのセグメント数が閾値未満になるように第2LRUの最もLRU側に存在するセグメントを再利用LRUリストのMRU側にリンクし直す処理を行い、それに合わせて第2LRUリスト情報742と再利用LRUリスト情報743を更新する。

【0116】ステップ2969で、ステップ2968において第2LRUリストから再利用LRUリストにリンクしなおされたセグメントに関して、キャッシュセグメント情報720中のステータス情報722を参照して、その値がダーティであるもののデータを物理記憶装置18に書き出す処理をディスクコントローラ制御部42に対して要求し、その完了を待つ。書き出し処理完了後、キャッシュセグメント情報720中の対応するステータス情報722をノーマルに変更し、ステップ2970に進む。

【0117】ステップ2970で処理を終了する。

【0118】図17に物理記憶位置管理・最適化部42が実施するデータ再配置処理の処理フローを示す。ここで、管理者の指示により処理を開始するモードと、あらかじめ設定されている時刻に自動的にデータ再配置案作成処理を実施し、その後に作成されたデータ再配置案を実現するためにデータ移動を自動実行するデータ再配置自動実行モードの2種類を考える。

【0119】後述するように、複数の異なった種類のデータ配置解析・データ再配置案作成処理を実行可能であり、処理すべき種類の指定をして処理を開始する。また、処理にパラメータが必要な場合は併せてそれが指定されているものとする。これらは、管理者が処理を指示する場合にはそのときに一緒に指示を出し、データ再配置自動実行モードの場合には処理する種類や必要なパラメータを事前に設定しておく。

【0120】ステップ2001でデータ再配置処理を開始する。このとき、データ配置解析・データ再配置案作成処理として何を実行するか指定する。また、必要であればパラメータを指定する。

【0121】ステップ2002でデータ再配置処理に必要なDBMSデータ情報36を前述した方法で取得し記憶する。なお、このデータ収集は、ステップ2001の処理開始とは無関係にあらかじめ実行しておくこともできる。この場合には、情報を取得した時点から現在まで情報に変更がないかどうかをこのステップで確認する。

【0122】ステップ2003では、ワーク領域を確保し、その初期化を行う。ワーク領域としては、図18に示すデータ再配置ワーク情報670と図19に示す移動プラン情報750を利用する。データ再配置ワーク情報670と移動プラン情報750の詳細とその初期データ作成方法は後述する。

【0123】ステップ2004でデータ配置の解析・再配置案の作成処理を実行する。後述するように、データ配置の解析・再配置案作成処理は複数の観点による異なったものが存在し、このステップではステップ2001で指定された処理を実行する。またステップ2001でパラメータを受け取った場合には、それを実行する処理に与える。

【0124】ステップ2005ではステップ2004のデータ再配置案作成処理が成功したかどうか確認する。成功した場合にはステップ2006に進む。失敗した場合にはステップ2010に進み、管理者にデータ再配置案作成が失敗したことを通知し、ステップ2011に進み処理を完了する。

【0125】ステップ2006では、現在データ再配置自動実行モードにより処理を実行しているか確認する。自動実行モードにより処理を実行している場合にはステップ2009に進む。そうでない場合には、ステップ2007に進む。

【0126】ステップ2007では、ステップ2004で作成されたデータ再配置案を管理者に提示する。この提示を受けて管理者はデータ再配置案に問題がないか判断する。ステップ2008では、データの再配置を続行するか否かを管理者から指示を受ける。続行する場合にはステップ2009に進み、そうでない場合にはステップ2011に進み処理を完了する。

【0127】ステップ2009では、ステップ2004で作成されたデータの再配置案を基にデータの再配置処理を実行する。このとき、移動プラン情報750中の移動順序761で示される順に指定されたボリュームの領域を指定された物理記憶装置18内の領域へデータの移動を実行する。移動処理機能の実現方法は前述した通りである。

【0128】ステップ2010でデータ再配置処理は完了である。

【0129】図18はステップ2003において作成する情報であるデータ再配置ワーク情報670を示す。データ再配置ワーク情報670はデータ移動可能領域を保持する空き領域情報680とデータ構造物理記憶位置情報712のコピーを保持する。空き領域情報680は、データ移動可能領域を示す物理記憶装置名502と物理ブロック番号514の組を保持する。

【0130】データの初期化は以下の方法で実行する。空き領域情報680はボリューム物理記憶位置メイン情報510中のボリューム名501が“Empty”であ

る領域を集めることにより初期化する。データ構造物理記憶位置情報712はDBMSデータ情報36に存在するデータをそのままコピーする。データ再配置案作成時にこれらのデータの値を変更するため、データ構造物理記憶位置情報712は必ずコピーを作成する。

【0131】図19はステップ2004で実行されるデータ配置解析・データ再配置案作成処理により作成されるデータ移動案を格納する移動プラン情報750を示す。移動プラン情報750は、移動指示の実行順序を示す移動順序761、移動するデータを持つボリュームとそのデータ領域を示す移動ボリューム名768と移動ボリューム論理ブロック番号769、そのデータの移動先の物理記憶装置とその記憶領域を示す移動先物理記憶装置名771と移動先物理ブロック番号772の組を保持する。この情報に関しては、何もデータを持たないように初期化する。

【0132】ステップ2004で実行されるデータ配置解析・データ再配置案作成処理について説明する。前述のように、この処理には幾つかの種類が存在する。全ての処理に共通するのは逐次的にデータ再配置のためのデータ移動案を作成することである。そのため、データ移動の順番には意味があり、移動プラン情報750中の移動順序761にその順番を保持し、その順序どおりにデータ移動を行うことによりデータの再配置を実施する。

【0133】また、逐次処理のため、移動後のデータ配置をもとに次のデータの移動方法を決定する必要がある。そこで、データ移動案を作成するたびにデータ再配置ワーク情報670をデータ移動後の配置に更新する。

【0134】データ再配置案作成時のデータ移動案の作成は以下のように行う。移動したいデータ量以上の連続した移動可能領域をデータ再配置ワーク情報670中の情報から把握し、その中の領域を適当に選択し、設定条件や後述する制約を満たすかどうか確認をする。もし、それらを満たす場合にはそこを移動先として設定する。それらを満たさない場合には他の領域を選択し、再度それらを満たすかどうか確認をする。

【0135】以下、設定条件と制約を満たす領域を発見するか、全ての移動したいデータ量以上の連続した移動可能領域が設定条件や制約を満たさないことを確認するまで処理を繰り返す。もし、全ての領域で設定条件や制約を満たさない場合にはデータ移動案の作成に失敗として終了する。

【0136】このときに重要なのは移動後のデータ配置において、問題となる配置を行わないことである。特にRDBMSにおいては、特定のデータに関してはアクセスが同時に行われる可能性が高く、それらを異なる物理記憶装置18上に配置する必要がある。そこで、以下で説明する全てのデータの移動案を作成する場合には、移動するデータに含まれるデータ構造と、移動先に含まれるデータ構造を調べ、ログとその他のデータ、一時表領

域とその他のデータ、表データとそれに対して作成された木構造の索引データがデータの移動後に同じ物理記憶装置18に配置されるかどうかを確認し、配置される場合には、その配置案は利用不可能と判断する。

【0137】なお、あるデータ構造がどの物理記憶装置18の領域に記憶されているか、また逆に、ある物理記憶装置18上の領域に記憶されるデータがどのデータ構造に対応するかは、データ再配置ワーク情報670中のデータ構造物理記憶位置情報712により把握可能である。

【0138】図20に第1のデータ配置解析・データ再配置案作成処理である、物理記憶装置稼働情報32を基にした同時アクセス実行データ構造を分離するためのデータ再配置案作成処理の処理フローを示す。本処理においては、物理記憶装置18の稼働率が閾値を超えたものはディスクネック状態にあると判断してそれを解消するデータの移動案を作成する。本処理は、実測値に基づいて問題点を把握し、それを解決する方法を見つけるため、より精度の高いデータ再配置案を作成すると考えられ、データ再配置自動実行モードで最も利用しやすいものである。

【0139】ステップ2101で処理を開始する。本処理を開始するにあたっては、どの期間の稼働率を参照するかを指定する。

【0140】ステップ2102では、物理記憶装置18の識別子と指定期間における物理記憶装置18の稼働率の組を記憶するワーク領域を取得し、物理記憶装置稼働情報32を参照してその情報を設定し、物理記憶装置18の稼働率で降順にソートする。物理記憶装置稼働情報32中では、同じ物理記憶装置18中に記憶されているデータであっても異なるボリュームのものは分離して稼働率を取得しているため、それらの総和として物理記憶装置18の稼働率を求める必要がある。

【0141】ステップ2103では、ステップ2102のソート結果をもとに物理記憶装置18の稼働率が閾値を超えているもののリストである過負荷確認リストを作成する。このリスト中のエントリに関しても稼働率が降順になるような順序を保つようにする。

【0142】ステップ2104では、過負荷確認リスト中にエントリが存在するか確認する。エントリが存在しない場合には、もう過負荷状態の物理記憶装置18が存在しないものとしてステップ2105に進みデータ再配置案作成処理成功として処理を終了する。エントリが存在する場合には、ステップ2106に進む。

【0143】ステップ2106では、過負荷確認リスト中の最も物理記憶装置18の稼働率が高いものを再配置対象の物理記憶装置18として選択する。

【0144】ステップ2107では、再配置対象となった物理記憶装置18内部のボリュームとその稼働率のリストを物理記憶装置稼働情報32を参照して作成し、稼

働率で降順にソートする。

【0145】ステップ2108では、リスト中のあるボリュームの稼働率があらかじめ定められた閾値を超えているかどうかを確認する。全てのボリュームの稼働率が閾値を超えていない場合には、ステップ2113に進み、あるボリュームの稼働率がその閾値を超えている場合には、ステップ2109に進む。

【0146】ステップ2109においては、稼働率が閾値を超えているボリュームに関して、確認対象の物理記憶装置18中に同時にアクセスされる可能性があるデータの組、すなわち、ログとその他のデータ、一時表領域とその他のデータ、表データとそれに対して作成された木構造の索引データがあるそのボリューム内部に記憶されているかどうかを発見する処理を行う。

【0147】ステップ2110では、ステップ2109における結果を確認し、同時アクセスデータ構造の組が存在する場合にはステップ2111に進む。同時アクセスデータ構造の組が存在しない場合には、ステップ2112に進む。

【0148】ステップ2111においては、同時アクセスデータ構造の組に属するデータを異なる物理記憶装置18に記憶するためのデータ移動案を作成し、ステップ2114に進む。

【0149】ステップ2112においては、現在確認対象となっているボリューム内のデータを論理ブロック番号に従って2分割し、その片方を他の物理記憶装置18へ移動するデータ移動案を作成し、ステップ2114に進む。

【0150】ステップ2113においては、現在確認対象になっている物理記憶装置18の稼働率が閾値を下回るまで、稼働率が高いボリュームから順に、その物理記憶装置18に記憶されているボリュームを構成するデータ全体を他の物理記憶装置18に移動するデータ移動案を作成し、ステップ2114に進む。

【0151】ステップ2111、2112、2113のデータ移動先を発見する際に、移動後の移動先の記憶装置の稼働率を予測する。物理記憶装置18毎の性能差が既知の場合にはその補正を行った移動データを含む記憶装置18上のボリュームの稼働率分、未知の場合には補正を行わない移動データを含む記憶装置18上のボリュームの稼働率分、データ移動により移動先の物理記憶装置18の稼働率が上昇すると考え、加算後の値が閾値を越えないような場所へのデータの移動案を作成する。

【0152】稼働率の加算分に関して、移動データ量の比率を考慮しても良いが、ここではデータ中のアクセスの偏りを考慮して移動データに全てのアクセスが集中したと考えた判断を行う。

【0153】ステップ2114では、データ移動案の作成に成功したかどうかを確認し、失敗した場合にはステップ2117に進みデータの再配置案作成処理失敗とし

て処理を終了する。成功した場合にはステップ2115に進む。

【0154】ステップ2115では作成したデータ移動案を移動プラン情報750に追加し、ステップ2116に進む。ステップ2116ではデータ再配置ワーク情報670を作成したデータ移動案に従って修正し、移動先記憶装置18のステップ2102で作成した物理記憶装置18毎の稼働情報の値を前述の移動後の稼働率判断値に修正する。その後、現在の確認対象の物理記憶装置18を過負荷確認リストから削除し、ステップ2104に戻り次の確認を行う。

【0155】次に第2のデータ配置解析・データ再配置案作成処理である、クエリ実行同時アクセスデータ構造カウント情報714を用いた同時アクセス実行データ構造を分離するためのデータ再配置案作成処理の処理フローを示す。本処理においては、クエリ実行同時アクセスデータ構造カウント情報714から同時にアクセスされるデータの組を取得し、それらを異なる物理記憶装置18に配置するデータ再配置案を作成する。

【0156】図21にクエリ実行時同時アクセスデータ構造カウント情報714を利用する同時アクセス実行データ構造を分離するためのデータ再配置案作成処理フローを示す。ステップ2201で処理を開始する。ステップ2202において、カウント値703の全エントリの総和に対してカウント値703の値が一定割合以上のデータ構造とその所属するDBMS110a, 110bの組を求め、それらを確認リストとして記憶する。

【0157】ステップ2203でステップ2202で求めた確認リスト中に含まれるデータ構造の組に関して、それらを異なる物理記憶装置18に記憶するデータ再配置案を作成し、ステップ2204に進む。なお、ステップ2203の処理に関しては、図22を用いて後で説明する。ステップ2204では、ステップ2203においてデータ再配置案の作成に成功したかどうかを確認し、成功した場合にはステップ2205に進みデータ再配置案作成処理成功として処理を終了し、失敗した場合にはステップ2206に進みデータ再配置案作成処理失敗として処理を終了する。

【0158】図22に指定されたデータ構造とそのデータ構造と同時にアクセスされる可能性が高いデータ構造の組を分離するデータ再配置案を作成する処理のフローを示す。本処理を開始するときには、データ構造名と物理記憶装置18から分離するデータ構造名の組のリストである確認リストを与える。

【0159】ステップ2301で処理を開始する。ステップ2303で確認リスト中にエントリが存在するか確認し、存在しない場合にはステップ2304に進みデータ再配置案作成処理成功として処理を終了する。存在する場合にはステップ2305に進む。

【0160】ステップ2305においては、確認リスト

から1つ確認対象データ構造名とその所属DBMS名の組とその分離データ構造名とその所属DBMS名の組の組を取得し、ステップ2306に進む。

【0161】ステップ2306においては、確認対象データ構造とその分離するデータ構造が同一の物理記憶装置上に記憶されているかどうかの確認を行う。この確認はデータ再配置ワーク情報670中のデータ構造物理記憶位置情報712を参照することにより可能である。両データ構造が全て異なる物理記憶装置上に存在する場合にはステップ2312に進み、ある物理記憶装置上に両データ構造が存在する場合にはステップ2307に進む。

【0162】ステップ2307においては、同一の物理記憶装置上に両データ構造が存在する部分に関してそれを分離するデータ移動案を作成する。ステップ2308においては、そのデータ移動案作成が成功したかどうか確認し、成功した場合にはステップ2310に進み、失敗した場合にはステップ2309に進みデータ再配置案作成処理失敗として処理を終了する。

【0163】ステップ2310においては、作成されたデータ移動案を移動プラン情報750に記憶する。ステップ2311においては、作成されたデータ移動案に従ってデータ再配置ワーク情報670を更新し、ステップ2312に進む。

【0164】ステップ2312においては、確認リストから現在確認対象となっているデータ構造の組のエントリを削除し、2303に進む。

【0165】図23に第3のデータ配置解析・データ再配置案作成処理である、データ構造の定義を基にした同時アクセス実行データ構造を分離するためのデータ再配置案作成処理の処理フローを示す。本処理においては、同時にアクセスされる可能性が高い、ログとその他のデータ、一時表領域とその他のデータ、表データとそれに対して作成された木構造の索引データが同一物理記憶装置18上に記憶されている部分が存在しないか確認をし、そのような部分が存在する場合にはそれを解決するデータ再配置案を作成する。

【0166】ステップ2401で処理を開始する。ステップ2402では、DBMSデータ構造情報621を参照して全てのログであるデータ構造名561とそれを利用するDBMS110a, 110bのDBMS名631の組を取得する。そして、そのデータ構造名とログ以外のデータを分離するデータ構造とする確認リストを作成し、ステップ2403に進む。ステップ2403ではステップ2402で作成した確認リストを用いてステップ2301から開始されるデータ構造分離のためのデータ再配置案作成処理を実行する。

【0167】ステップ2404ではステップ2403におけるデータ再配置案作成処理が成功したか確認をし、成功した場合にはステップ2405に進む。失敗した場

合にはステップ2412に進みデータ再配置案作成処理失敗として処理を終了する。

【0168】ステップ2405では、DBMSデータ構造情報621を参照して全ての一時表領域であるデータ構造名561とそれを利用するDBMS110a、110bのDBMS名631の組を取得する。そして、そのデータ構造と一時表領域以外のデータを分離するデータ構造とする確認リストを作成し、ステップ2406に進む。ステップ2406ではステップ2405で作成した確認リストを用いてステップ2301から開始されるデータ構造分離のためのデータ再配置案作成処理を実行する。

【0169】ステップ2407ではステップ2406におけるデータ再配置案作成処理が成功したか確認をし、成功した場合にはステップ2408に進む。失敗した場合にはステップ2412に進みデータ再配置案作成処理失敗として処理を終了する。

【0170】ステップ2408では、DBMS索引定義情報624を参照して全ての木構造索引の索引名635とそれを利用するDBMS110a、110bのDBMS名631の組とそれに対応する表のデータ構造名とそれを利用するDBMS110a、110bのDBMS名631の組を対応表情報637から取得する。そして、それらの索引と表に関するデータを組とする確認リストを作成し、ステップ2409に進む。ステップ2409ではステップ2408で作成した確認リストを用いてステップ2301から開始されるデータ構造分離のためのデータ再配置案作成処理を実行する。ステップ2410ではステップ2409におけるデータ再配置案作成処理が成功したか確認をし、成功した場合にはステップ2411に進み、データ再配置案作成処理成功として処理を終了する。失敗した場合にはステップ2412に進みデータ再配置案作成処理失敗として処理を終了する。

【0171】図24に第4のデータ配置解析・データ再配置案作成処理である、特定の表や索引の同一データ構造に対するアクセス並列度を考慮したデータ再配置案作成処理の処理フローを示す。この処理は、ランダムアクセス実行時の処理の並列度を考慮してディスクネックの軽減を図るためにデータの再配置を行うものである。この処理を実行する際には、データ再配置の確認対象とするデータ構造をDBMS名631とデータ構造名561の組として指定する。

【0172】ステップ2501で処理を開始する。ステップ2502において、指定されたデータ構造の物理記憶装置上に割り当てられた記憶領域利用総量を求める。この値は、DBMSデータ構造情報621中のデータ構造データ量641を参照することにより求める。

【0173】ステップ2503においては、DBMSデータ構造情報621を参照して指定データ構造における最大アクセス並列度569を取得する。

【0174】ステップ2504において、ステップ2502で求めた指定データ構造の記憶領域利用総量をステップ2503で求めた最大アクセス並列度569で割った値を、指定データ構造の1つの物理記憶装置18上への割り当てを許可する最大量として求める。この制約により、特定の物理記憶装置18に偏ることなく最大アクセス並列度569以上の台数の物理記憶装置18に指定データ構造が分散して記憶されることになり、最大アクセス並列度569による並列度でランダムアクセスが実行されてもディスクネックになりにくい状況となる。なお、割り当て許可最大量の値は、実際のアクセス特性を考慮してこの方法で求めた値から更に増減させても構わない。

【0175】ステップ2505において、指定データ構造のデータがステップ2504で求めた最大量を超えて1つの物理記憶装置18上に割り当てられているものが存在するかデータ再配置ワーク情報670を用いて確認し、そのようなものが存在しない場合にはステップ2509に進み、データ再配置案作成処理成功として処理を終了する。存在する場合にはステップ2506に進む。

【0176】ステップ2506においては、ステップ2504で求めた最大量を超えて1つの物理記憶装置18上に割り当てられている部分を解消するデータ移動案を作成する。このとき、移動案作成に考慮するデータ移動量は指定データ構造の現在の物理記憶装置18上への割り当て量のステップ2504で求めた最大量からの超過分以上である必要がある。また、移動先物理記憶装置18においても、移動後にステップ2504で求めた最大量を超過しないようにする必要がある。

【0177】ステップ2507においては、ステップ2506のデータ移動案作成処理が成功したか確認をする。成功した場合にはステップ2508に進む。失敗した場合にはステップ2510に進み、データ再配置案作成処理失敗として処理を終了する。

【0178】ステップ2508においては作成したデータ移動案を移動プラン情報750に記憶し、ステップ2509に進みデータ再配置案作成処理成功として処理を終了する。

【0179】図25に第5のデータ配置解析・データ再配置案作成処理である、特定の表データに対するシーケンシャルアクセス時のディスクネックを解消するデータ再配置案作成処理の処理フローを示す。この処理を実行する際には、データ再配置の確認対象とする表をDBMS名631とデータ構造名561の組として指定する。

【0180】前述のように、対象とするDBMS110a、110bの種類が絞られるが、データ構造物理記憶位置情報712はシーケンシャルアクセス順にソートされてデータを記憶しているため、シーケンシャルアクセス方法は既知である。

【0181】また、並列にシーケンシャルアクセスを実



行する場合に、その領域の分割法は並列にアクセスしない場合のシーケンシャルにアクセスする順番を並列度に合わせて等分に分割するものとする。

【0182】この並列アクセスによる分割後の1つのアクセス領域を全て同一の物理記憶装置18上に配置するのは必ずしも現実的ではない。そこで、分割後のアクセス領域がある一定量以上連続にまとまって1つの物理記憶装置上に記憶されていればよいと判断する。ただし、どのような場合でも連続してアクセスされることがなく、分割後のアクセス領域が異なるものに分類されるものに関しては、並列シーケンシャルアクセス時にアクセスがぶつかる可能性があるため、異なる物理記憶装置18に記憶するという指針を設けて、これに沿うようなデータ配置を作成することによりシーケンシャルアクセスの性能を高める。

【0183】ステップ2601で処理を開始する。ステップ2602において、指定された表の物理記憶装置上に割り当てられた記憶領域利用総量を求める。この値は、DBMSデータ構造情報621中のデータ構造データ量641を参照することにより求める。ステップ2603においては、DBMSデータ構造情報621を参照して指定データ構造における最大アクセス並列度569を取得する。

【0184】ステップ2604において、ステップ2602で求めた指定表の記憶領域利用総量をステップ2603で求めた最大アクセス並列度569で割った量が、並列アクセス時にシーケンシャルにアクセスされる1つの領域のデータ量である。データ構造物理記憶位置情報712はシーケンシャルアクセス実行順にソートされているため、これを用いて最大アクセス並列度569の並列アクセスが実行されると仮定した前述のデータ分割指針を作成する。

【0185】ステップ2605において、データ再配置ワーク情報670を参照しながら、指定データ構造はステップ2604で作成したデータ分割指針に沿ったデータ配置が物理記憶装置18上で行われているか確認し、そうであればステップ2609に進み、データ再配置案作成処理成功として処理を終了する。そうでない場合にはステップ2606に進む。

【0186】ステップ2606においては、物理記憶装置18上において、ステップ2604で求めたデータ分割指針に従うデータ配置を求める。このとき、データがある一定値以下の領域に細分化されている場合には、連続した空き領域を探し、そこにアクセス構造を保つようにデータを移動するデータ移動案を作成する。また、最大アクセス並列度569の並列アクセスにより異なるアクセス領域に分離されるデータが同じ物理記憶装置18上に配置されないようなデータ移動案を作成する。

【0187】ステップ2607においては、ステップ2606のデータ移動案作成処理が成功したか確認をす

る。成功した場合にはステップ2608に進み、失敗した場合にはステップ2610に進み、データ再配置案作成処理失敗として処理を終了する。

【0188】ステップ2608においては作成したデータ移動案を移動プラン情報750に記憶し、ステップ2609に進みデータ再配置案作成処理成功として処理を終了する。＜第二の実施の形態＞本実施の形態では、DBMSが実行される計算機とファイルを管理単位とする記憶装置がネットワークを用いて接続された計算機システムにおいて、記憶装置がDBMSに関する情報、記憶装置外におけるデータの記憶位置のマッピングに関する情報を取得し、それらを用いて記憶装置の動作を改善する。

【0189】記憶装置において、記憶装置内部でデータの記憶位置を動的に変更する機能を有し、取得した情報をもとに好適なデータ再配置案を作成し、データの記憶位置の動的変更機能を用いて、作成したデータ再配置案に従ったデータ配置を実現し、アクセス性能を改善する。また、取得情報をもとにしたデータキャッシュの制御を行いより良いアクセス性能特性が得られるようにする。

【0190】図26は、本発明の第二の実施の形態における計算機システムの構成図である。図示されたように、本発明の第二の実施の形態は本発明の第一の実施の形態と以下の点が異なる。

【0191】本実施の形態においてはI/Oバスインターフェイス70、I/Oバス71、I/Oバススイッチ72が存在せず、記憶制御装置10bとDBホスト80c、80dはネットワーク79を介してのみ接続される。記憶装置10はファイルを単位としたデータ記憶管理を行う記憶装置10bに変更される。そのため、物理記憶装置稼働情報32、データキャッシュ管理情報34、DBMSデータ情報36、ボリューム物理記憶位置管理情報38がそれぞれ物理記憶装置稼働情報32b、データキャッシュ管理情報34b、DBMSデータ情報36b、ファイル記憶管理情報38bに変更される。

【0192】DBホスト80c、80dで実行されるOS100ではボリュームマネージャ102、ファイルシステム104が削除されその代わりに記憶装置10bが提供するファイルをアクセスするための機能を有するネットワークファイルシステム104bが追加され、OS100が保持するマッピング情報106がネットワークマウント情報106bへ変更される。

【0193】記憶装置10はファイルを管理単位とする記憶装置10bに変更される。DBホスト80c、80dからのアクセスもNFS等のファイルをベースとしたプロトコルで実施される。記憶装置10におけるボリュームの役割は、記憶装置10bにおいてはファイルもしくはファイルを管理するファイルシステムとなり、そのファイルの記憶位置管理情報がファイル記憶管理情報3

8bである。1つの記憶装置10bの中に複数のファイルシステムが存在しても構わない。物理記憶装置18の稼動情報はボリュームを単位とした取得からファイルシステムまたはファイルを単位とした取得に変更する。記憶装置10b内にファイルシステムが存在する場合でもデータの移動機能を実現可能である。

【0194】図27はDBホスト80c、80dのOS100内に記憶されているネットワークマウント情報106bを示す。ネットワークマウント情報106bは、記憶装置10bから提供され、DBホスト80c、80dにおいてマウントされているファイルシステムの情報で、ファイルシステムの提供元記憶装置とそのファイルシステムの識別子である記憶装置名583とファイルシステム名1001、そして、そのファイルシステムのマウントポイントの情報であるマウントポイント1031の組を保持する。

【0195】図28は記憶装置10b内に保持されるファイル記憶管理情報38bを示す。図5のボリューム物理記憶位置管理情報38からの変更点は、ボリューム物理記憶位置メイン情報510、ボリュームデータ移動管理情報511からファイル物理記憶位置情報510b、ファイルデータ移動管理情報511bにそれぞれ変更される。上記の変更内容は、ボリュームの識別子であるボリューム名501がファイルの識別子となるファイルシステム名1001とファイルパス名1002に、ボリューム内のデータ領域を示すボリューム論理ブロック番号512と移動論理ブロック番号782がそれぞれファイルブロック番号1003または移動ファイルブロック番号1021に変更されるものである。

【0196】ここで、ファイルパス名1002が“Empty”であるエントリ1015は特殊なエントリであり、このエントリには記憶装置10b内の物理記憶装置18の領域のうち、指定ファイルシステム内でファイルの記憶領域として割り当てられていない領域を示し、図5中のボリュームデータ移動管理情報511を用いるデータ移動方式で説明した処理手順を用い、この領域に対して移動するデータをコピーすることによりデータの物理記憶位置の動的変更機能を実現する。

【0197】ここで注意が必要なのは、データ移動案作成時にデータ移動先の制約が増えた点である。本実施の形態においては、ファイルシステムを複数保持することが許されている。一般のファイルシステムにおいては、あるファイルシステムが他のファイルシステムが管理する領域を利用することは不可能である。つまり、一般のファイルシステムを用いている場合には、ファイルの移動は、そのファイルが存在しているファイルシステム内に閉じる必要がある。ただし、あるファイルシステムが他のファイルシステムが管理する領域を利用可能な機構を有している場合にはこの限りではない。

【0198】図29に記憶装置10b内に保持される物

理記憶装置稼動情報32bを示す。図6の物理記憶装置稼動情報32からの変更点は、稼動情報取得単位がボリュームからファイルシステムに変更されたため、ボリューム名501の部分がファイルシステム名1001に変更されたことである。また、稼動情報取得単位をファイルとしてもよく、このときはボリューム名501の部分がファイルシステム名1001とファイルパス名1002に変更される。

【0199】図30に記憶装置10b内に保持されているDBMSデータ情報36bを示す。図7のDBMSデータ情報36からの変更点は、ボリュームを利用した記憶管理からファイル利用した記憶管理に変更されたためデータ構造物理記憶位置情報712に修正が加えられ、データ構造物理記憶位置情報712bに変更されたことである。

【0200】図31にDBMSデータ情報36b中に含まれるデータ構造物理記憶位置情報712bを示す。図9のデータ構造物理記憶位置情報712からの変更点は、ボリュームを利用した記憶管理からファイル利用した記憶管理に変更されたため、ボリューム名501とボリューム論理ブロック番号512の部分がファイルシステム名1001とファイルパス名1002とファイルブロック番号1003に変更されたことである。この情報は、DBMSデータ記憶位置情報622とネットワークマウント情報106bを記憶装置10の外部から取得し、さらにファイル物理記憶位置情報510bを参照して、対応する部分を組み合わせることにより作成する。

【0201】図32に記憶装置10b内に保持されているデータキャッシュ管理情報34bを示す。図13のデータキャッシュ管理情報34からの変更点は、ボリュームを利用した記憶管理からファイル利用した記憶管理に変更されたため、キャッシュセグメント情報720に修正が加えられ、キャッシュセグメント情報720bに変更されたことである。キャッシュセグメント情報720bのキャッシュセグメント情報720からの変更点は、上述の理由により、ボリューム名501とボリューム論理ブロック番号512の部分がファイルシステム名1001とファイルパス名1002とファイルブロック番号1003に変更されたことである。

【0202】図33にステップ2003において作成する情報であるデータ再配置ワーク情報670bを示す。図18のデータ再配置ワーク情報670からの変更点は、ボリュームを利用した記憶管理からファイル利用した記憶管理に変更されたため、空き領域情報680とデータ構造物理記憶位置情報712に修正が加えられ、それぞれ空き領域情報680bとデータ構造物理記憶位置情報712bへ変更されたことである。空き領域情報680bの空き領域情報680からの変更点は、ファイルシステムを利用した領域管理を実施しているため、空き領域管理はファイルシステムを意識する必要があり、空

き領域情報としては、データの記憶に利用していない場所を示す物理記憶装置名 502 と物理ブロック番号 514 とその空き領域を管理するファイルシステム名 1001 の組を保持する。空き領域情報 680b はファイル物理記憶位置メイン情報 510b 中のファイルパス名 1001 が “Empty” である領域を集めることにより初期化する。

【0203】図 34 はステップ 2004 で実行されるデータ配置解析・データ再配置案作成処理により作成されるデータ移動案を格納する移動プラン情報 750b を示す。図 19 の移動プラン情報 750 からの変更点は、ボリュームを利用した記憶管理からファイル利用した記憶管理に変更されたため、移動ボリューム名 568 と移動ボリューム論理ブロック番号 769 の部分が移動ファイルシステム名 1101 と移動ファイルパス名 1102 と移動ファイルブロック番号 1103 に変更されたことである。

【0204】前述のように、記憶装置 10 において一般のファイルシステムを用いている場合には、ファイルの移動は、そのファイルが存在しているファイルシステム内に閉じる必要がある。従って、データ再配置案作成処理に関して、本実施の形態における第一の実施の形態から変更点は、データ移動先は現在データが存在しているファイルシステム上に限られるという制約が追加される。ただし、この制約も記憶装置 10 が利用しているファイルシステムが他のファイルシステムが管理する領域を利用可能な機構を有している場合には除かれる。

【0205】記憶装置 10b における本実施の形態における第一の実施の形態からの差は、ほとんどがボリューム名 501 をファイルシステム名 1001 とファイルパス名 1002 に、ボリューム論理ブロック番号 512 をファイルブロック番号 1003 に変更することであり、その他の変更点もその差を述べてきた。記憶装置 10b における処理に関しては、前記のデータ再配置案作成処理における制約を除き、基本的にこれまで述べてきた変更点と同じ変更点への対応方法を実施すれば、第一の実施の形態における処理をほぼそのまま本実施の形態に当てはめることができる。

【0206】

【発明の効果】本発明により以下のことが可能となる。第一に、DBMS が管理するデータを保持する記憶装置において、DBMS の処理の特徴を考慮することにより DBMS に対してより好ましい性能特性を持つことができる。この記憶装置を用いることにより、既存の DBMS に対してプログラムの修正無しに DBMS 稼働システムの性能を向上させることができるようになる。つまり、高性能な DB システムを容易に構築できるようになる。

【0207】第二に、記憶装置の性能最適化機能を提供するため、それにより記憶装置の性能に関する管理コス

トを削減することができる。特に、本発明は、DB システムの高性能化に寄与するため、この記憶装置を用いた DB システムの性能に関する管理コストを削減することができる。更に、本発明を用いた記憶装置は、自動で DBMS の特性を考慮したデータ配置の改善を行うことができ、管理コストの削減に大きく寄与する。

【図面の簡単な説明】

【図 1】第一の実施の形態における計算機システムの構成を示す図である。

【図 2】DB ホスト 80a, 80b の OS 100 内に記憶されているマッピング情報 106 を示す図である。

【図 3】DBMS 110a, 110b 内に記憶されているその内部で定義・管理しているデータその他の管理情報であるスキーマ情報 114 を示す図である。

【図 4】DB ホスト 80a, 80b のメモリ 88 上に記憶されている実行履歴情報 122 を示す図である。

【図 5】記憶装置 10 内に保持されているボリューム物理記憶位置管理情報 38 を示す図である。

【図 6】記憶装置 10 内に保持されている物理記憶装置稼働情報 32 を示す図である。

【図 7】記憶装置 10 内に保持されている DBMS データ情報 36 を示す図である。

【図 8】DBMS データ情報 36 中に含まれる DBMS スキーマ情報 711 を示す図である。

【図 9】DBMS データ情報 36 中に含まれるデータ構造物理記憶位置情報 712 を示す図である。

【図 10】DBMS データ情報 36 中に含まれるクエリ実行同時アクセスデータ構造カウント情報 714 を示す図である。

【図 11】DBMS データ情報 36 に含まれる DBMS データ構造キャッシュ効果情報 715 を示す図である。

【図 12】記憶装置 10 において指定されたデータ構造のデータをデータキャッシュに保持する効果があるかどうかの判断する処理のフローを示す図である。

【図 13】記憶装置 10 内に保持されているデータキャッシュ管理情報 34 を示す図である。

【図 14】記憶装置 10 がホストからデータの読出し要求を受け取ったときの処理フローを示す図である。

【図 15】記憶装置 10 がホストからデータの書き込み要求を受け取ったときの処理フローを示す図である。

【図 16】アクセス先のデータの内容に従いアクセス要求のあったデータを保持するセグメントを適当な管理リストに繋ぐ処理のフローを示す図である。

【図 17】記憶装置 10 内で実施されるデータ再配置処理の処理フローを示す図である。

【図 18】データ配置解析・再配置案作成処理で利用するデータ再配置ワーク情報 670 を示す図である。

【図 19】データ配置解析・再配置案作成処理で作成されるデータ移動案を格納する移動プラン情報 750 を示す図である。

【図20】物理記憶装置稼動情報32を基にした同時アクセス実行データ構造を分離するためのデータ再配置案作成処理の処理フローを示す図である。

【図21】クエリ実行時同時アクセスデータカウント情報714を利用する同時アクセス実行データ構造を分離するためのデータ再配置案作成処理の処理フローを示す図である。

【図22】指定されたデータ構造とそのデータ構造と同時にアクセスされる可能性が高いデータ構造の組を分離するデータ再配置案を作成する処理のフローを示す図である。

【図23】データ構造の定義を基にした同時アクセス実行データ構造を分離するためのデータ再配置案作成処理の処理フローを示す図である。

【図24】特定の表や索引の同一データ構造に対するアクセス並列度を考慮したデータ再配置案作成処理の処理フローを示す図である。

【図25】特定の表データに対するシーケンシャルアクセス時のディスクネックを解消するデータ再配置案作成処理の処理フローを示す図である。

【図26】第二の実施の形態における計算機システムの構成を示す図である。

【図27】DBホスト80c、80dのOS100内に記憶されているネットワークマウント情報106bを示す図である。

【図28】記憶装置10b内に保持されるファイル記憶管理情報38bを示す図である。

【図29】記憶装置10b内に保持される物理記憶装置稼動情報32bを示す図である。

【図30】記憶装置10b内に保持されているDBMSデータ情報36bを示す図である。

【図31】DBMSデータ情報36b中に含まれるデータ構造物理記憶位置情報712bを示す図である。

【図32】記憶装置10b内に保持されているデータキャッシュ管理情報34bを示す図である。

【図33】データ配置解析・再配置案作成処理で利用するデータ再配置ワーク情報670bを示す図である。

【図34】データ配置解析・再配置案作成処理で作成されるデータ移動案を格納する移動プラン情報750bを示す図である。

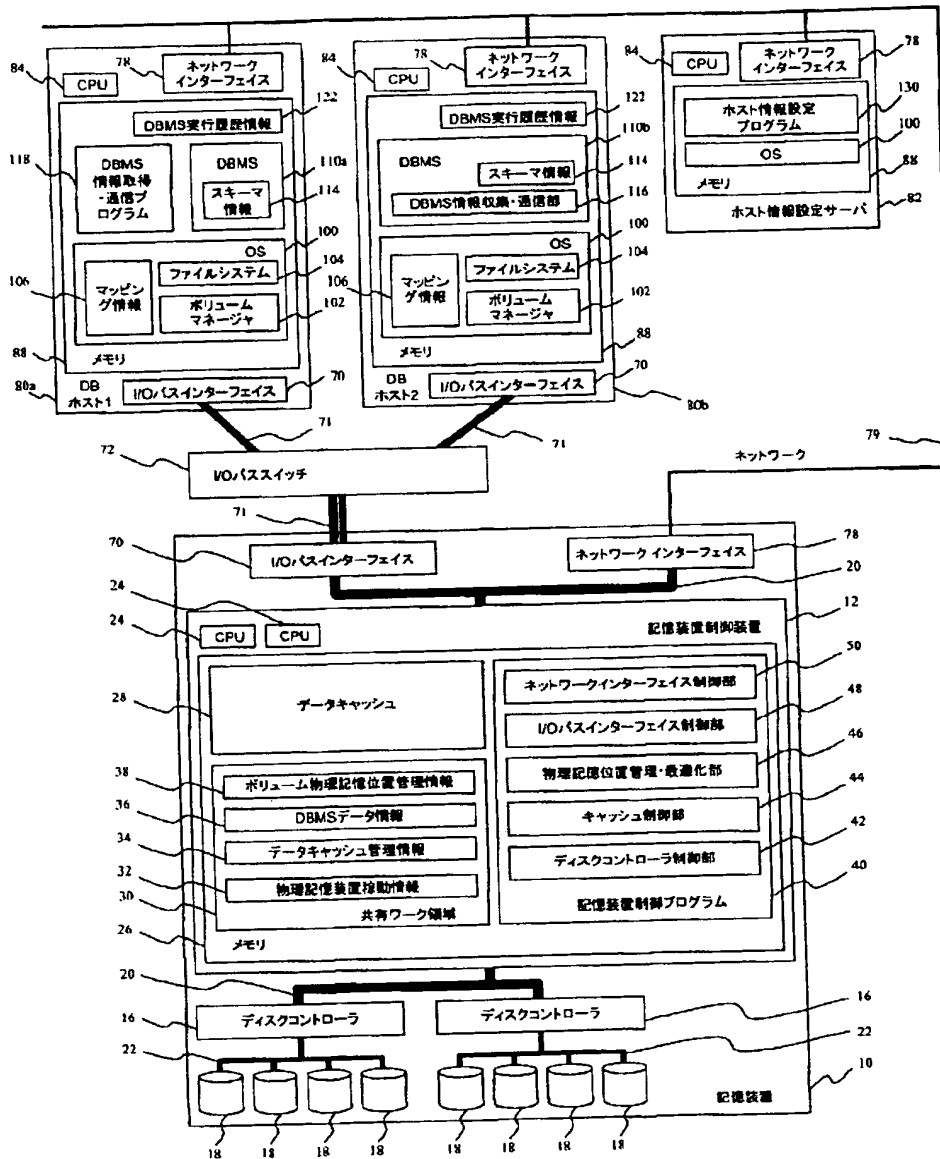
【符号の説明】

10、10b 記憶装置  
18 物理記憶装置

28  
32、32b  
34、34b  
36、36b  
38  
理情報  
38b  
40  
42  
部  
44  
46  
部  
48  
ス制御部  
50  
イス制御部  
70  
ス  
71  
72  
78  
イス  
79  
80a、80b、80c、80d DBホスト  
82  
100  
ステム)  
102  
104  
104b  
テム  
106  
106b  
110a、110b  
理システム)  
114  
116  
118  
ログラム  
122  
130  
データキャッシュ  
物理記憶装置稼動情報  
データキャッシュ管理情報  
DBMSデータ情報  
ボリューム物理記憶位置管  
ファイル記憶管理情報  
記憶装置制御プログラム  
ディスクコントローラ制御  
キャッシュ制御部  
物理記憶位置管理・最適化  
I/Oバスインターフェイ  
ネットワークインターフェ  
I/Oバスインターフェイ  
I/Oバス  
I/Oバススイッチ  
ネットワークインターフェ  
ネットワーク  
80a、80b、80c、80d DBホスト  
ホスト情報設定サーバ  
OS（オペレーティングシ  
ボリュームマネージャ  
ファイルシステム  
ネットワークファイルシス  
マッピング情報  
ネットワークマウント情報  
DBMS（データベース管  
スキーマ情報  
DBMS情報通信部  
DBMS情報取得・通信ブ  
実行履歴情報  
ホスト情報設定プログラム

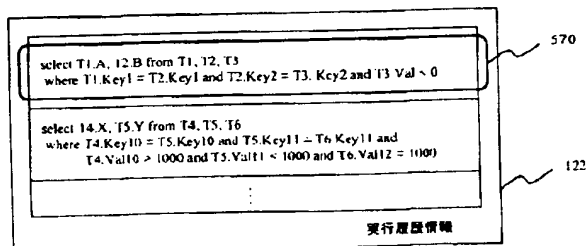
【図1】

図1



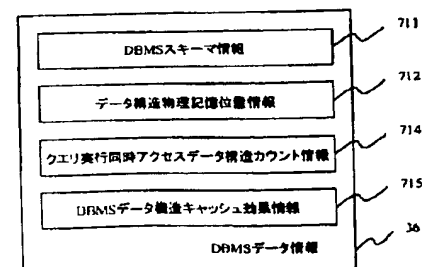
【図4】

図4



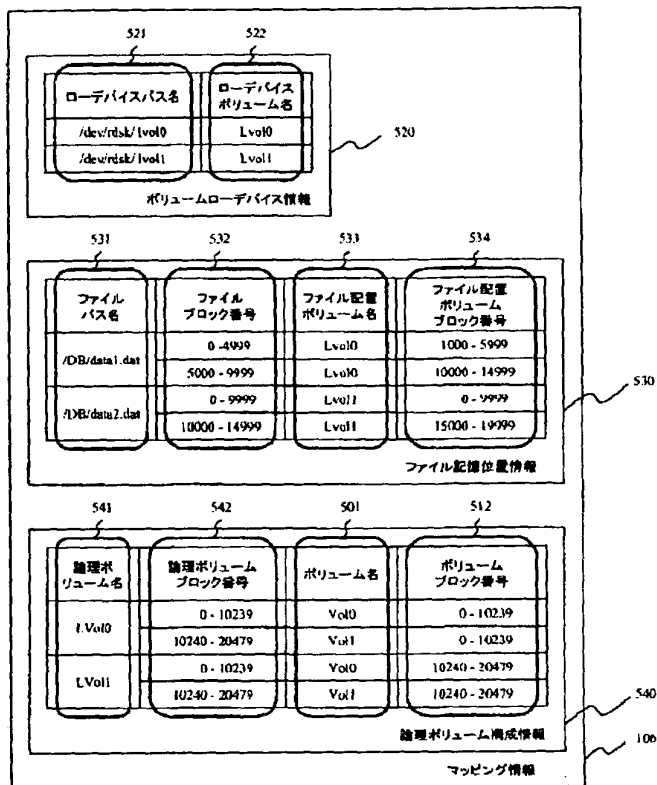
【図7】

図7



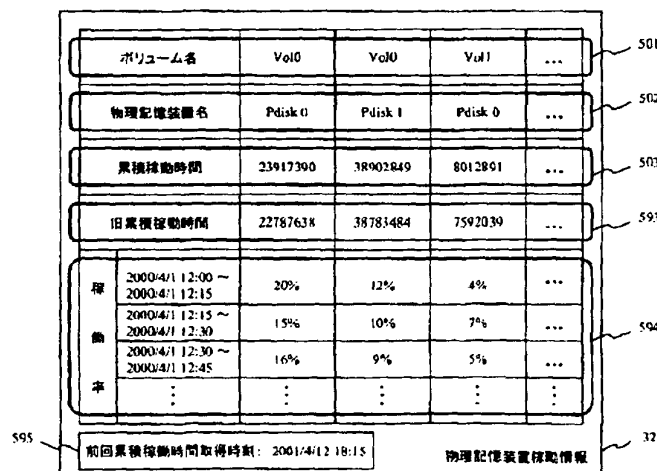
【図2】

図2



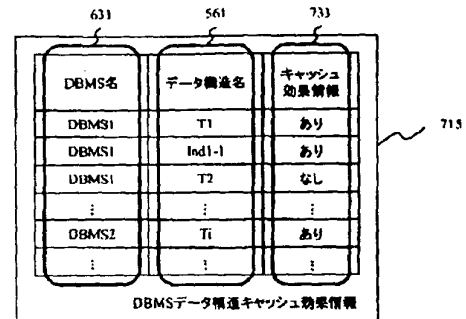
【図6】

図6



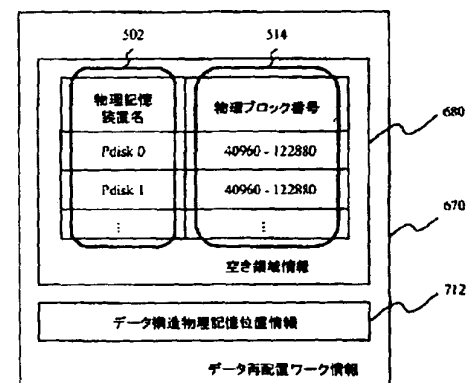
【図11】

図11



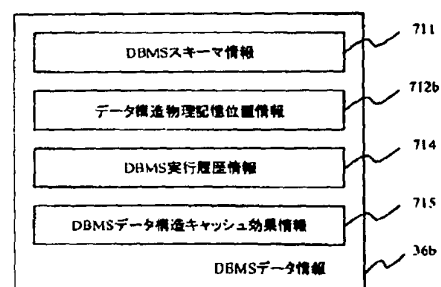
【図18】

図18

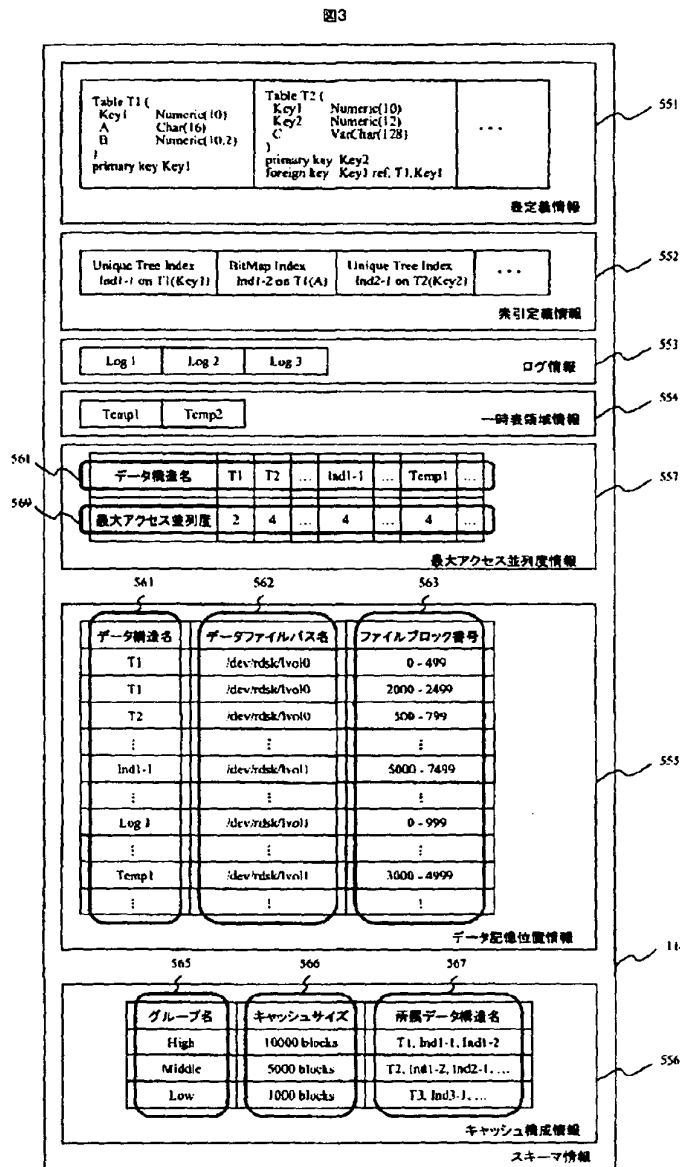


【図30】

図30

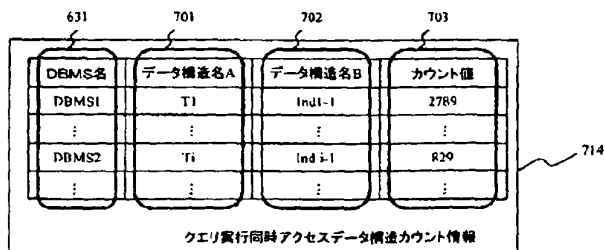


【図3】



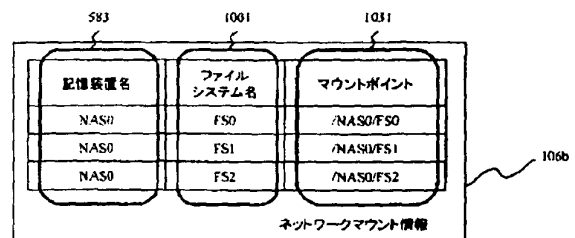
【図10】

図10



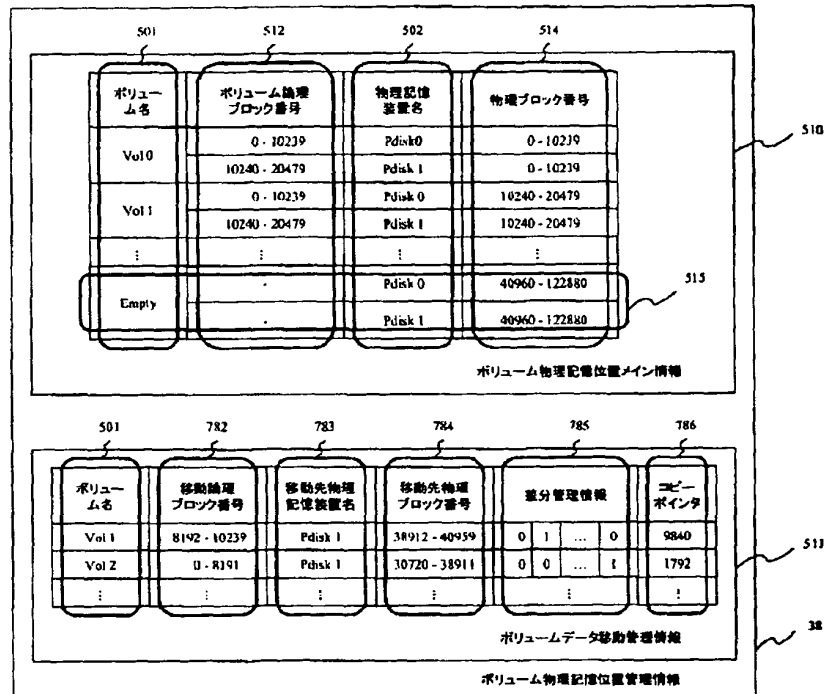
【図27】

図27



【図5】

図5



【図9】

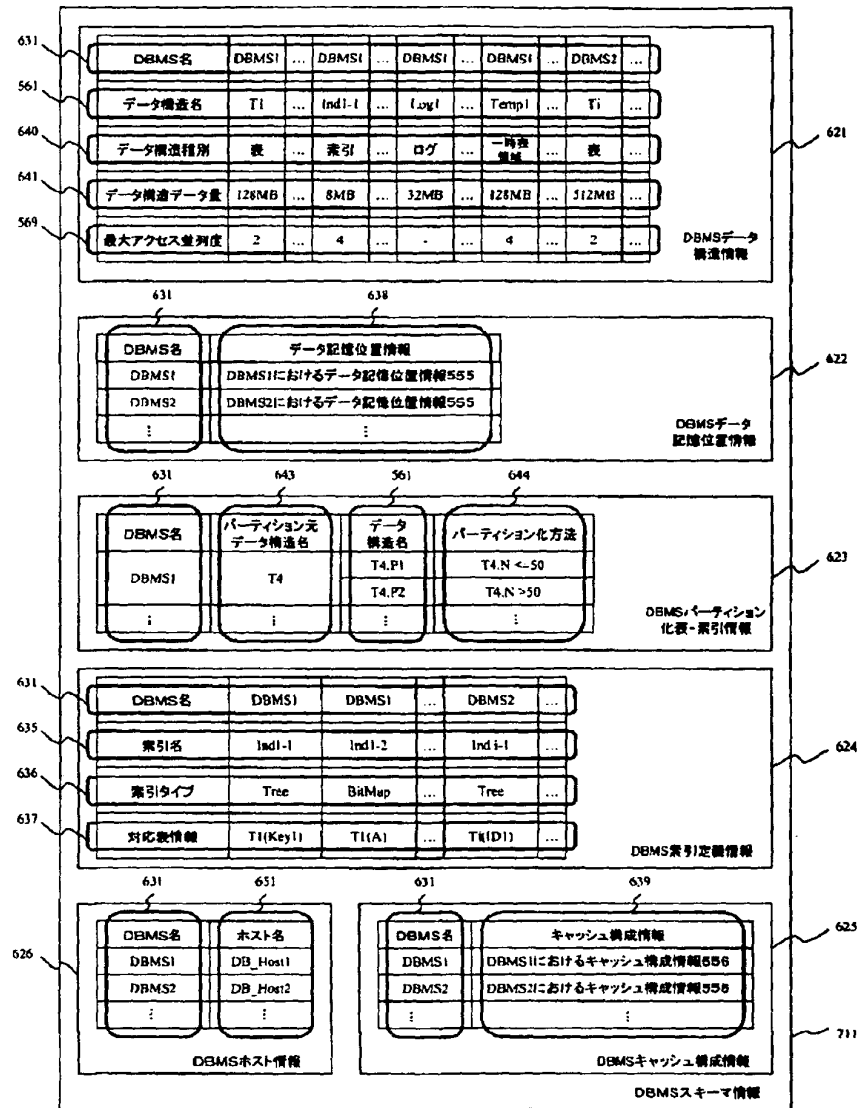
図9

DBMS名	DBMS1	DBMS1	DBMS1	...	DBMS2	...
データ構造名	T1	T1	T2	...	T1	...
ボリューム名	Vol0	Vol1	Vol0	...	Vol2	...
ボリュームブロック番号	0 - 4999	5000 - 9999	0 - 219	...	0 - 9999	...
物理記憶装置名	Pdisk0	Pdisk0	Pdisk0	...	Pdisk1	...
物理ブロック番号	0 - 4999	10240 - 15239	10080 - 10239	...	20480 - 30479	...



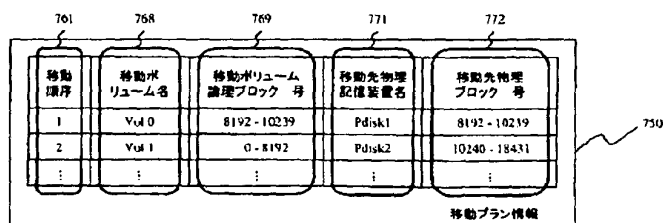
【図8】

図8



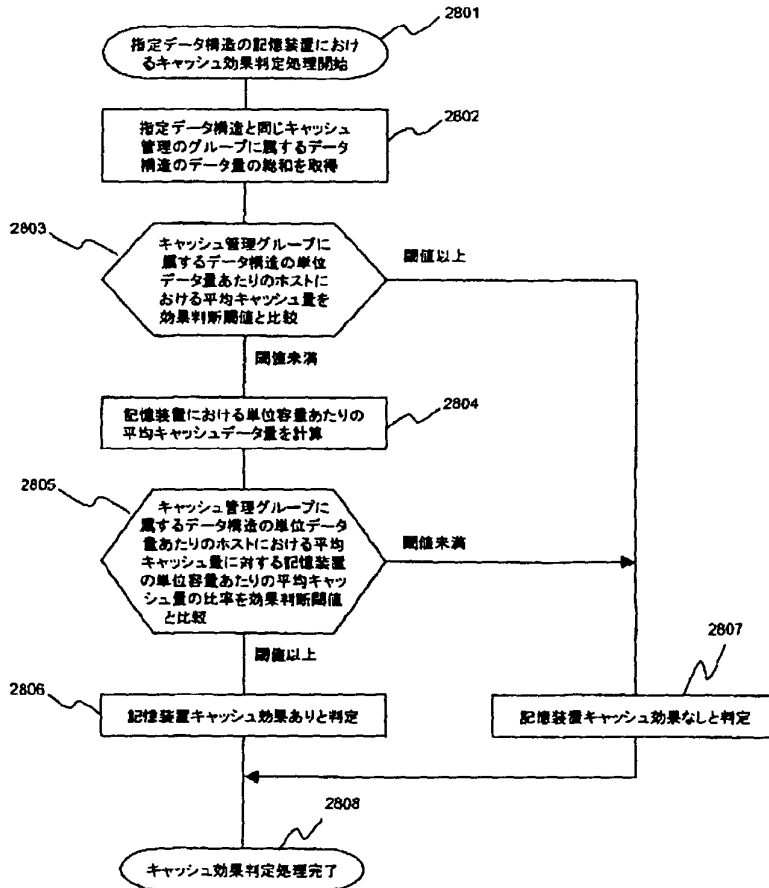
【図19】

図19



【図12】

図12



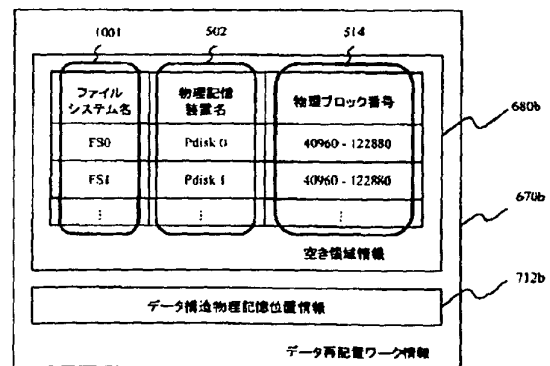
【図29】

図29

ファイルシステム名						1001
		FS0	FS0	FS1	...	
物理記憶装置名						502
		Pdisk 0	Pdisk 1	Pdisk 0	...	
累積移動時間						592
		23911390	38902849	8012891	...	
旧累積移動時間						593
		22787638	38783484	7592039	...	
検 査 時 間	2000/4/1 12:00 ~ 2000/4/1 12:15	20%	12%	4%	...	594
	2000/4/1 12:15 ~ 2000/4/1 12:30	15%	10%	7%	...	
	2000/4/1 12:30 ~ 2000/4/1 12:45	16%	9%	5%	...	
	...	...	...	...	...	
	...	...	...	...	...	
前回累積移動時間取得時刻: 2001/4/2 18:15						595
物理記憶装置稼働情報						32b

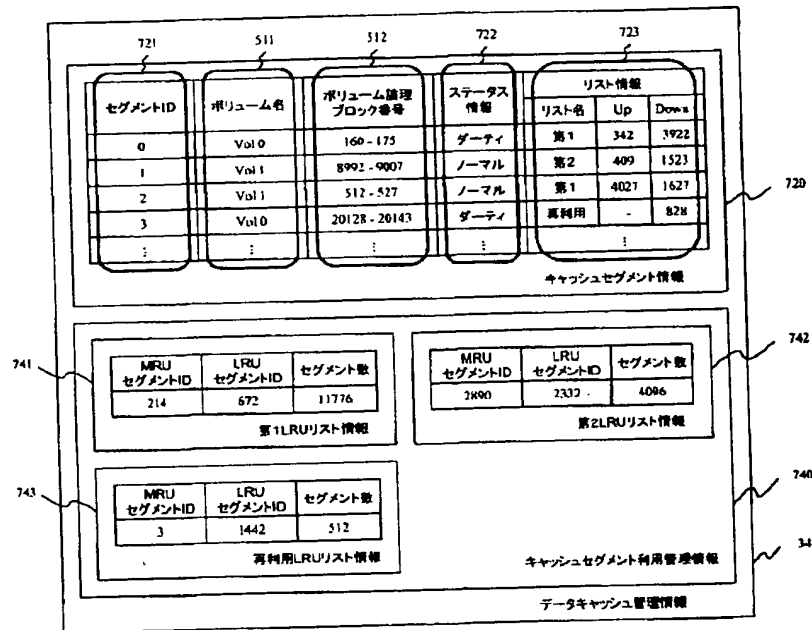
【図33】

図33



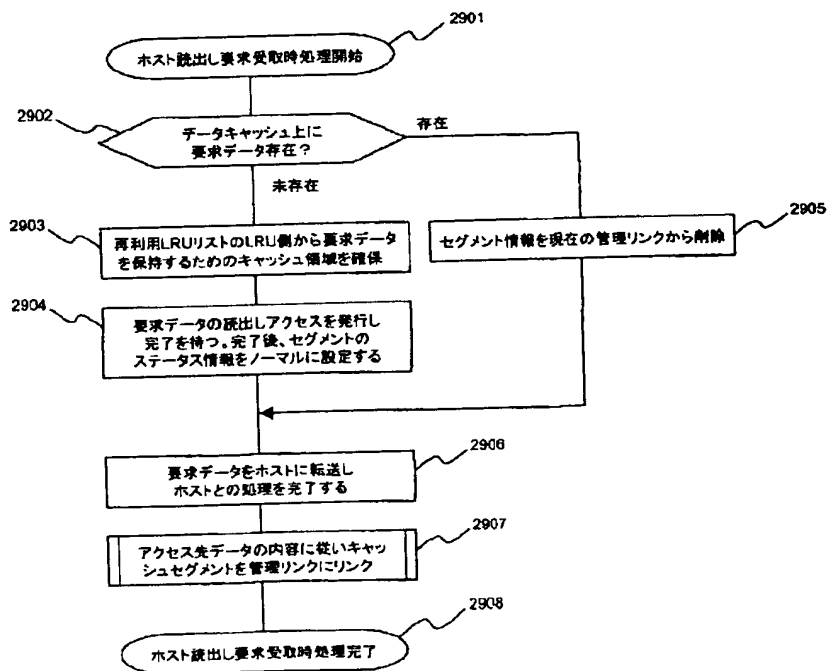
【図13】

図13



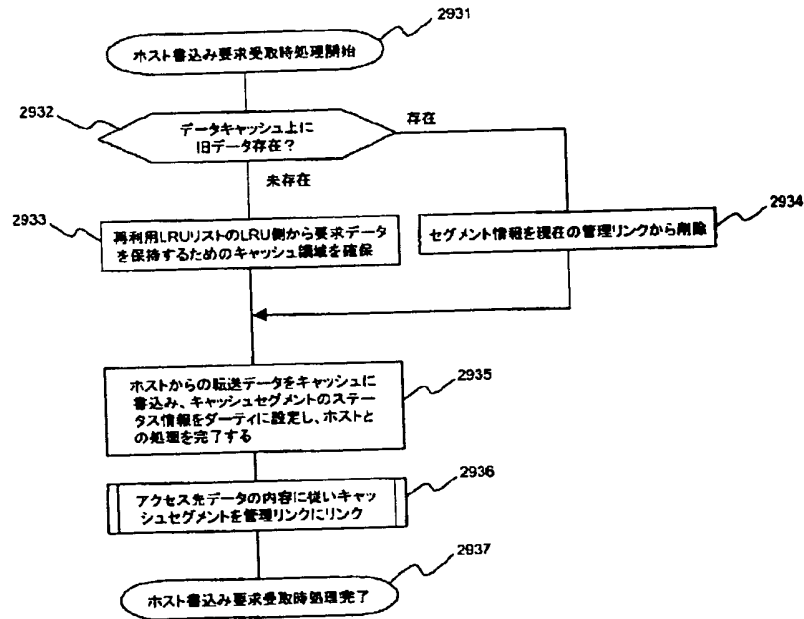
【図14】

図14



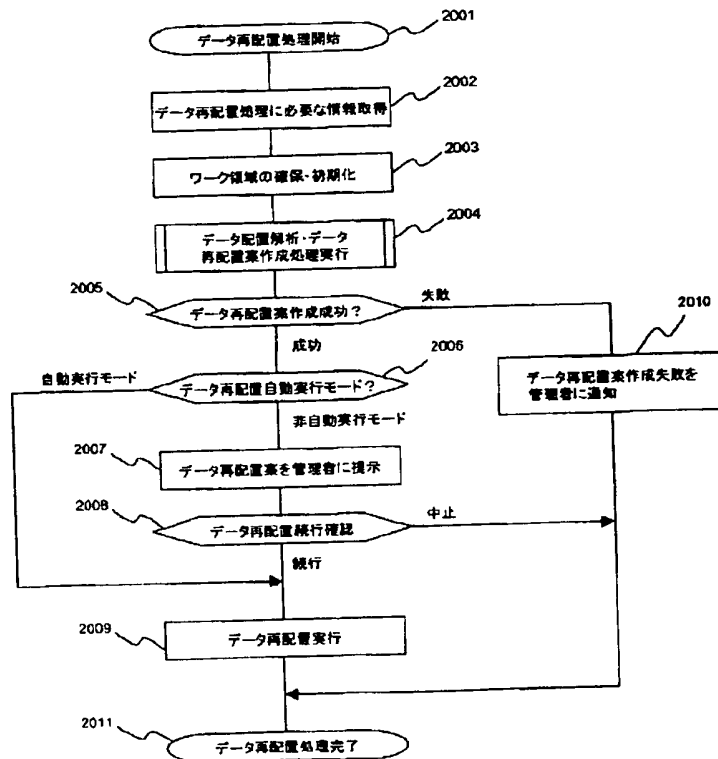
【図15】

図15

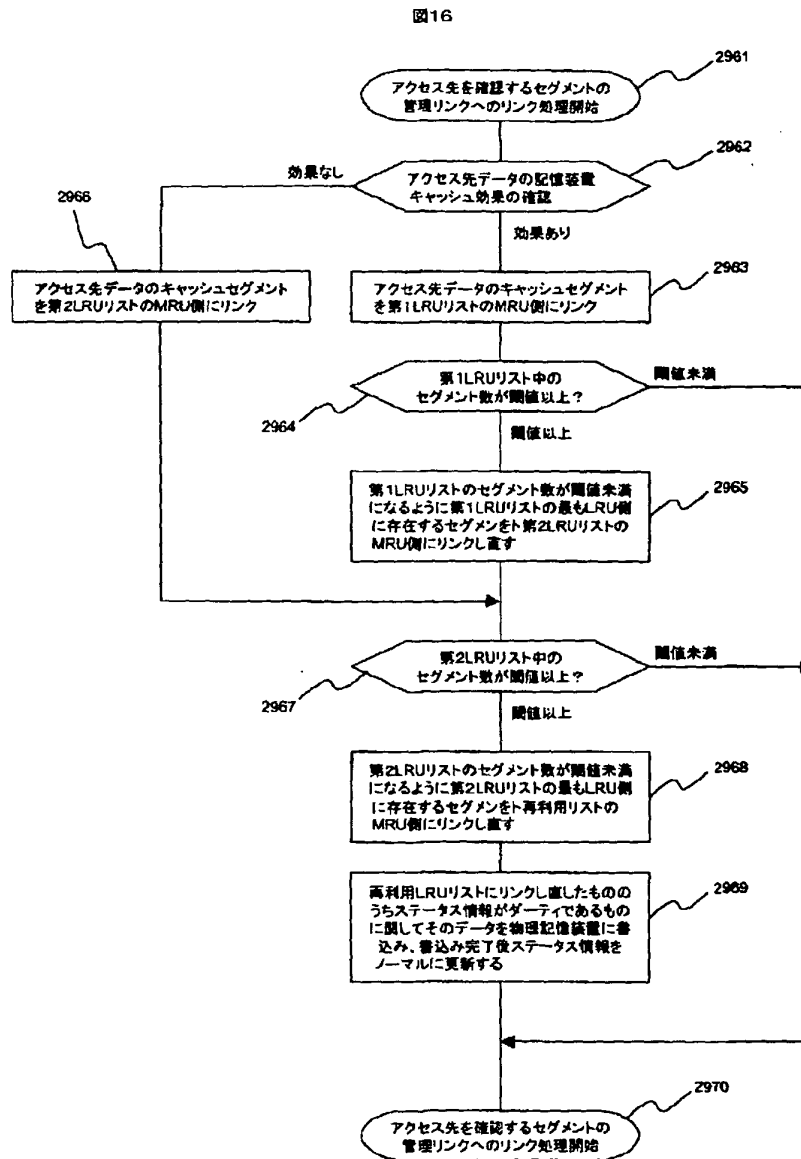


【図17】

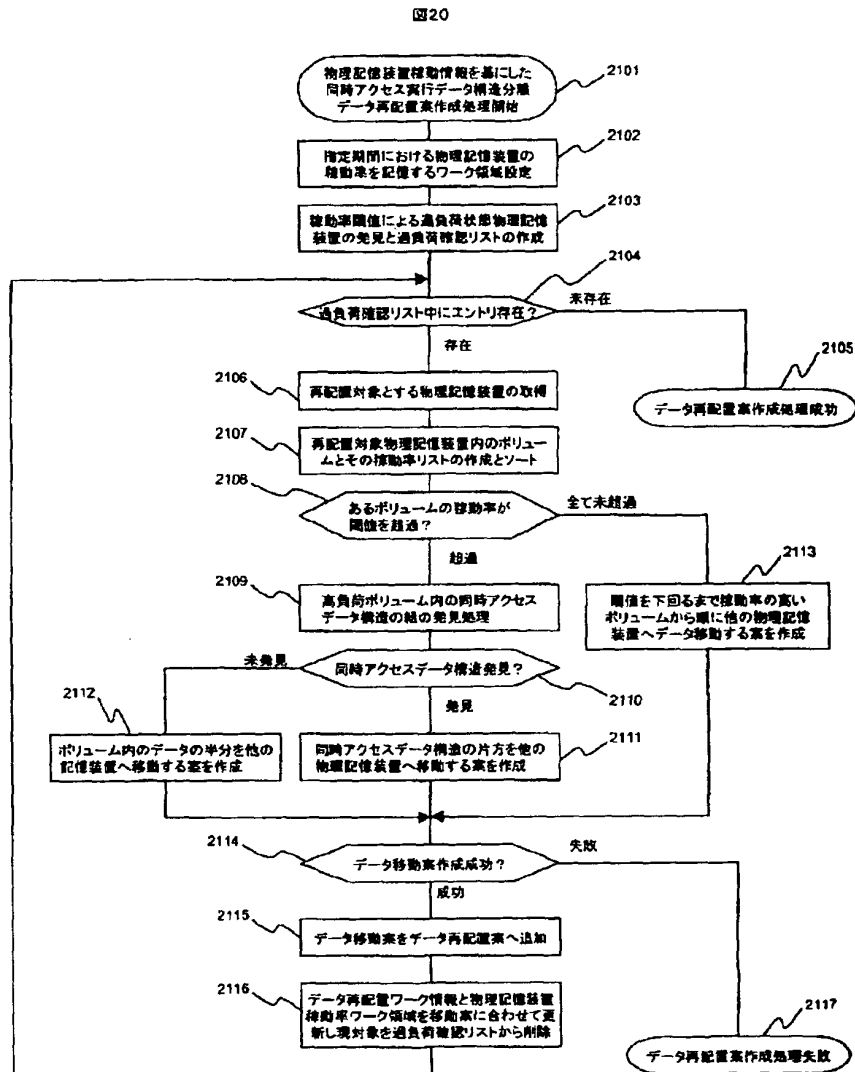
図17



【図16】



【図20】



【図34】

図34

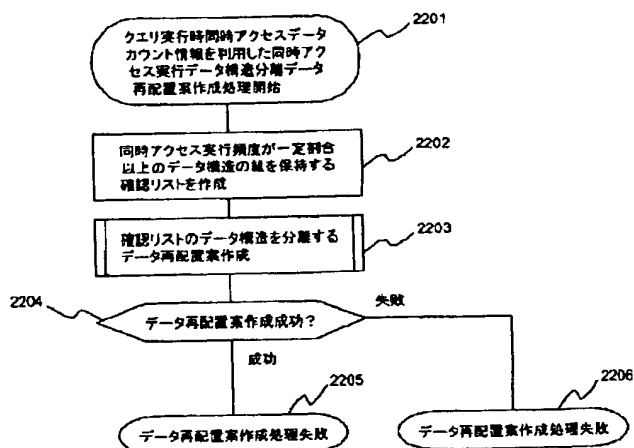
761	1101	1102	1103	171	772
移動 順序	移動ファイル システム名	移動ファイル パス名	移動ファイル ブロック番号	移動先物理 記憶装置名	移動先物理 ブロック番号
1	FS 0	/control.dat	0 - 1023	Pdisk1	8192 - 9213
2	FS 2	/data2.dat	0 - 4095	Pdisk3	10240 - 14335
⋮	⋮	⋮	⋮	⋮	⋮

移動プラン情報

750b

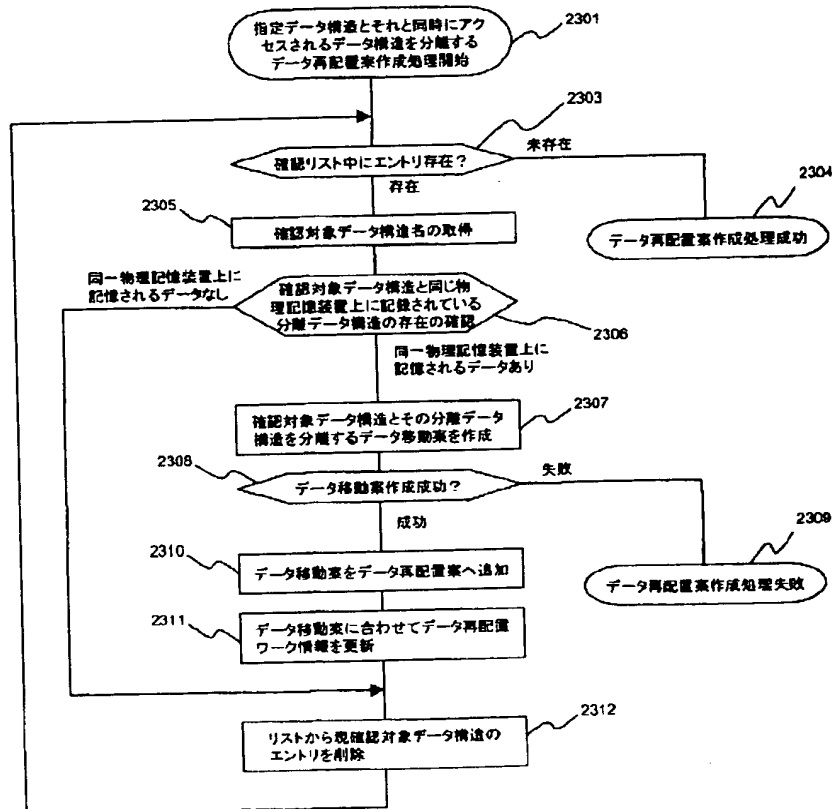
【図21】

図21



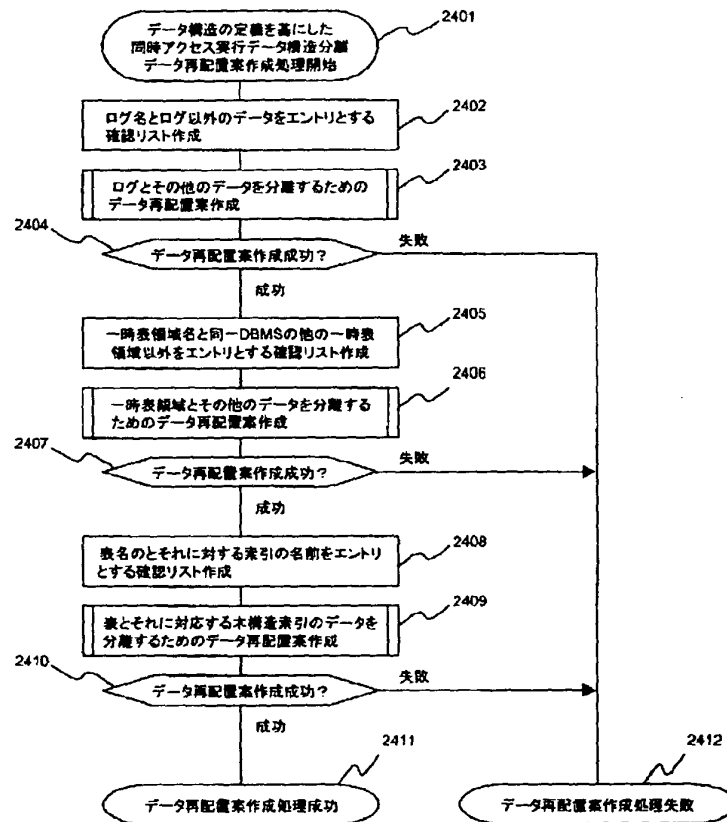
【図22】

図22



【図23】

図23



【図31】

図31

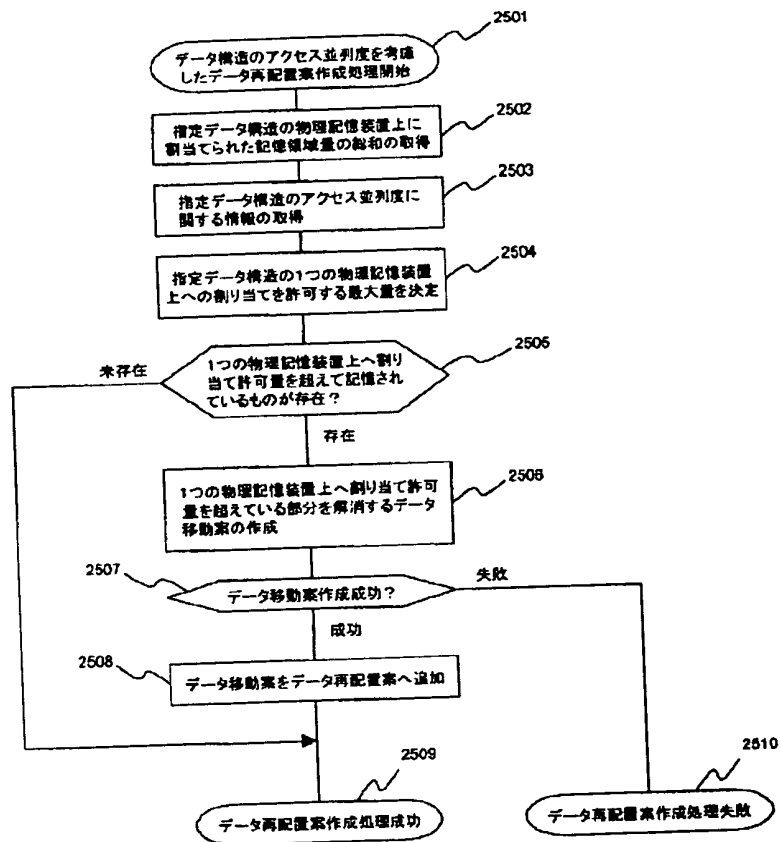
DBMS名	DBMS1	DBMS1	DBMS1	...	DBMS2	...
データ構造名	T1	T1	T2	...	Ti	...
ファイルシステム名	FS1	FS1	FS1	...	FS2	...
ファイルパス名	/data1.dat	/data1.dat	/data2.dat	...	/datai.dat	...
ファイルブロック番号	0 - 4999	5000 - 9999	0 - 4999	...	0 - 9999	...
物理記憶装置名	Pdisk 1	Pdisk 0	Pdisk 1	...	Pdisk 2	...
物理ブロック番号	10240 - 15239	10240 - 15239	15240 - 20239	...	20480 - 30479	...

データ構造物理記憶位置情報



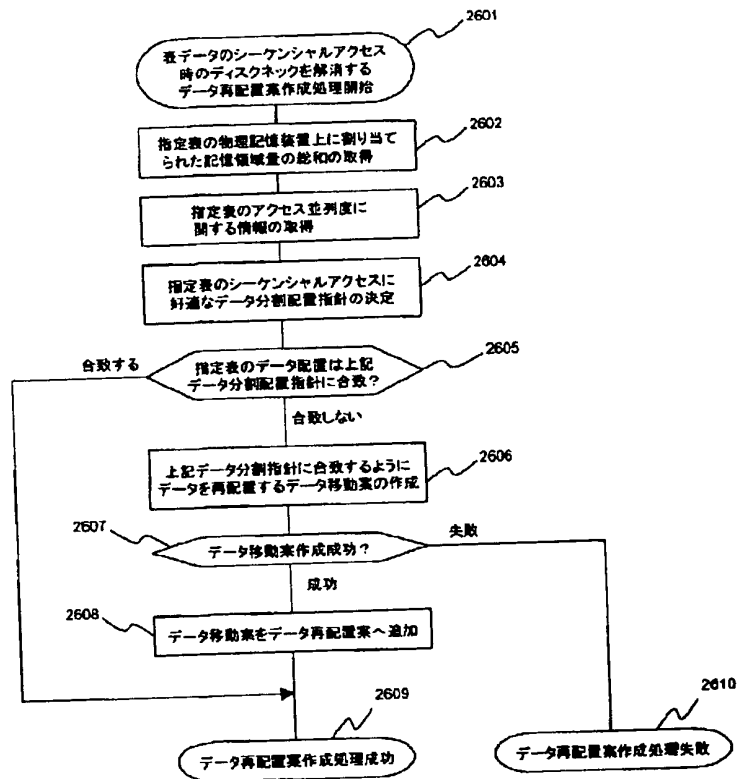
【図24】

図24



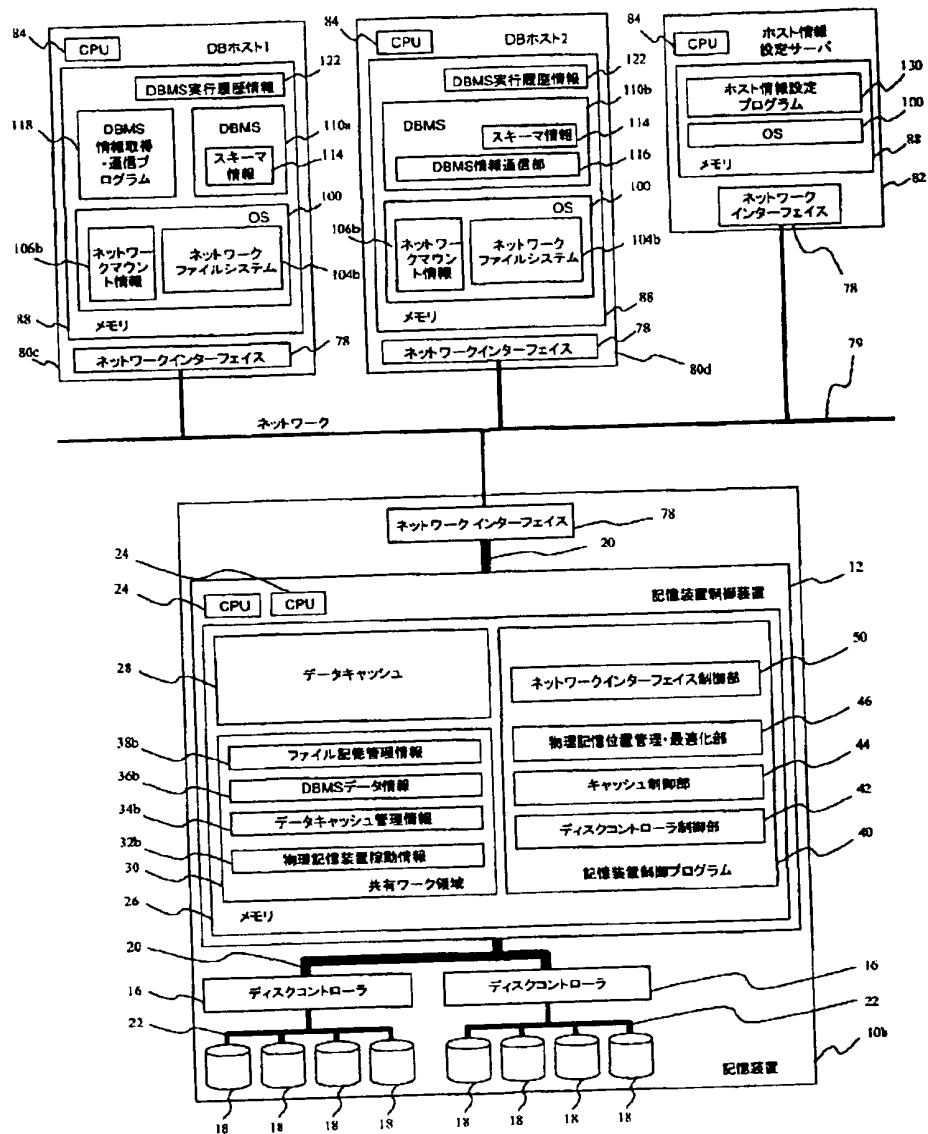
【図25】

図25



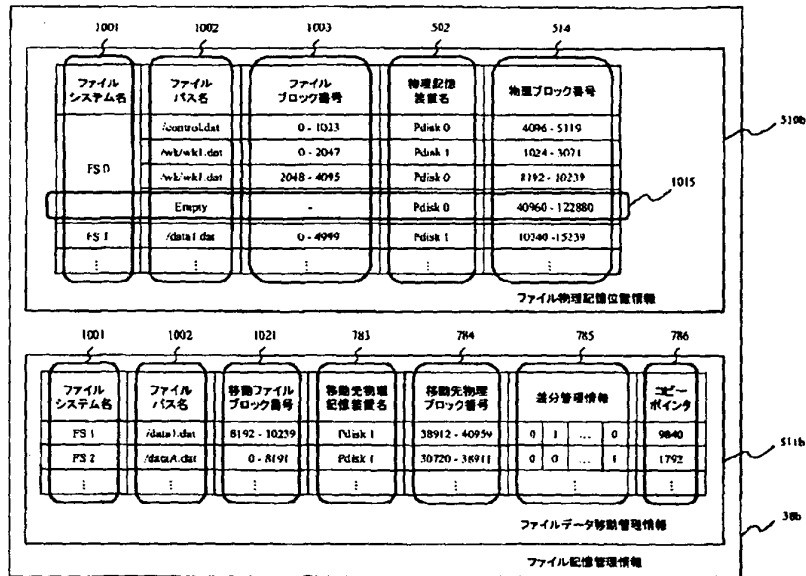
【図26】

図26



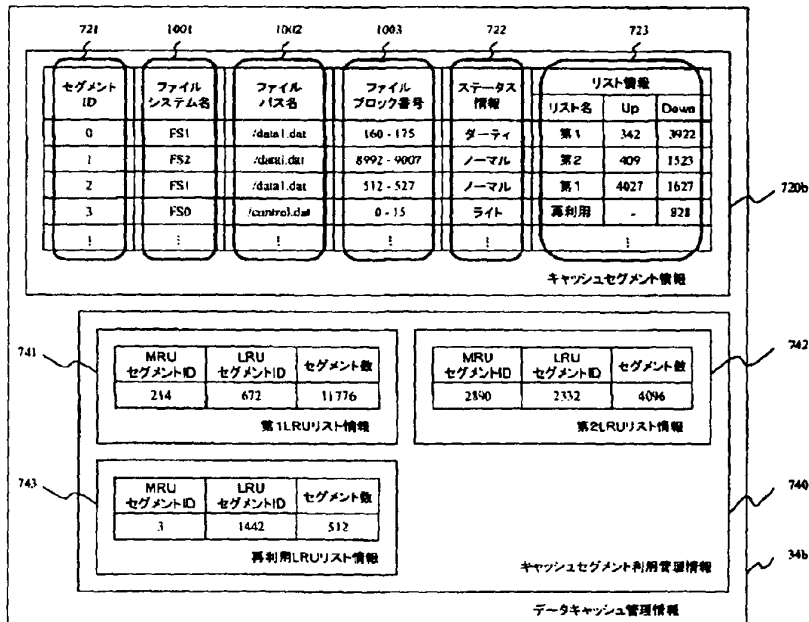
【図28】

図28



【図32】

図32



フロントページの続き

(72)発明者 喜連川 優  
千葉県松戸市二十世紀が丘丸山町17

F ターム(参考) 5B005 JJ13 LL17 MM04 QQ02 QQ04  
VV02  
5B065 BA01 CA11 CC02 CC08 CE11  
CH01 CH18  
5B082 BA09 FA12